# Data Analytics

# Table of Content

[**CHAPTER 1: INTRODUCTION TO STATISTICS**](#_heading=h.w7jj6yi6jlht) **10**

[Theory](#_heading=h.t1r9roulivsb) 10

[Statistics – Big Picture](#_heading=h.ny2aoh92czp3) 10

[Basic Statistical Terminology](#_heading=h.imylnzfub2b6) 11

[Parameter and Statistic](#_heading=h.nv1u40aifso9) 11

[Variables and Data](#_heading=h.t3yhhrpt1p5a) 12

[Variables – Dependent and Independent](#_heading=h.g4xitzfnagy6) 12

[Data](#_heading=h.kshgo37k0mgu) 14

[Levels of Measurement Scales](#_heading=h.o4krhn11b1gh) 14

[Descriptive Statistics](#_heading=h.3143wroeuvwi) 15

[What is a Correlation Matrix?](#_heading=h.dyzsgfg4qek2) 24

[AIM](#_heading=h.5xxv0w1flqqn) 26

[Introduction to Python and Python versions](#_heading=h.y43ezla05mv6) 27

[Anaconda Python Distribution](#_heading=h.jh3s5hxf6jir) 28

[Why Anaconda?](#_heading=h.486740mhpgq4) 28

[Anaconda Installation](#_heading=h.9ak2c3d0exzl) 29

[What are Jupyter notebooks?](#_heading=h.itqk4abc01qw) 33

[How does a notebook work?](#_heading=h.lhth8xt6nabh) 33

[Installing and Launching the Jupyter notebook](#_heading=h.j4a5ab8jwhox) 34

[Launching Jupyter notebook](#_heading=h.aa3w6qqvrgtj) 34

[Importing the necessary packages](#_heading=h.l1ntp6diyqxo) 35

[Calculating Central Tendencies](#_heading=h.ynfc55zgwxvw) 36

[Calculating Measure of Dispersion](#_heading=h.o74uobq3enzm) 36

[Task 2: Standard Deviation](#_heading=h.1hyep9nh992v) 37

[Task 3: Variation](#_heading=h.gvb3ivcgjre4) 38

[Task 4: Range and Percentile](#_heading=h.8vow94og9yv) 38

[Range](#_heading=h.74jkrawy2yn2) 38

[Percentile](#_heading=h.egw8tp2wgh3r) 39

[Task 5: Quartiles](#_heading=h.m8oi9lxb323a) 39

[Task 6: Boxplot](#_heading=h.61kmqrclco7k) 40

[Task 7: Skewness](#_heading=h.2c7qu73ylw9u) 42

[Negatively skewed distribution](#_heading=h.s1g3l7usr4f5) 43

[Positively skewed distribution](#_heading=h.49wskcv38jfg) 44

[Case Study](#_heading=h.y067k1c9sfa9) 44

[Descriptive Statistics](#_heading=h.do0a6dcc2fqf) 47

[Assessment](#_heading=h.dt9w59e8a1j4) 51

[Choose the appropriate option](#_heading=h.xzdlrdgb5ts9) 51

[Fill in the spaces with appropriate answers](#_heading=h.is3edgtrof3k) 53

[Programming Assignment](#_heading=h.f5wlqewl6ybk) 53

[Solutions for Assessment](#_heading=h.64np8u2o7635) 55

[Choose the appropriate options answers](#_heading=h.53skinyo496s) 55

[Fill in the spaces with appropriate answers:](#_heading=h.iq4g32h2qr5j) 55

[Programming Assignment Solution](#_heading=h.xhq19hcypxwv) 55

[**CHAPTER 2: PROBABILITY DISTRIBUTIONS**](#_heading=h.rv03rawki1mu) **57**

[Theory](#_heading=h.1iu570ct3jhx) 57

[Assigning Probabilities](#_heading=h.awjk9xfl5rx2) 57

[Probability Terminology](#_heading=h.9ysk65xlupp9) 60

[Probability – Rules](#_heading=h.w6p8xf5q10ft) 60

[Probability – Types](#_heading=h.bo60im4j1d66) 62

[Probability Distributions](#_heading=h.vxy842ge041) 66

[Types of Discrete Probability Distributions](#_heading=h.wzuab14khsrp) 69

[Poisson distribution Vs Binomial distribution](#_heading=h.la3mdos3yu6g) 74

[AIM](#_heading=h.kc791q66cbez) 75

[Importing the necessary packages](#_heading=h.2jctls4i2ubm) 77

[Calculating probability using binomial distribution](#_heading=h.324z8iren7td) 77

[Calculating the probability using Poisson distribution](#_heading=h.p5ik1o72u4i8) 82

[Case Study](#_heading=h.p9ec32t7m77g) 84

[Assessment](#_heading=h.3ipto37teh05) 86

[Choose the appropriate option](#_heading=h.9uodfqflk2jl) 86

[Fill in the spaces with appropriate answers](#_heading=h.avr1bf3vgc8p) 87

[Programming Assignment](#_heading=h.9lrfjjq2u6uu) 87

[Solutions for Assessment](#_heading=h.wgspyzk9001a) 89

[Choose the appropriate options answers](#_heading=h.sqm45mm41pbe) 89

[Fill in the spaces with appropriate answers](#_heading=h.ntqxws1x0rnp) 89

[Programming Assignment Solutions](#_heading=h.2tgcx6rr26k3) 89

[**CHAPTER 3 – INFERENTIAL STATISTICS**](#_heading=h.s31uf5pf2iws) **90**

[Theory](#_heading=h.4v9i2459g7q6) 90

[Continuous Random Variable](#_heading=h.ljnbbxu8lvwx) 90

[Types of Continuous Random Variables](#_heading=h.o1dbxusubnr8) 91

[Sampling Distribution](#_heading=h.3bl7goup0vgb) 102

[Sampling Distribution of the Means](#_heading=h.sy6qwajpxb4e) 102

[Expectation and Standard deviation of 𝑋 bar](#_heading=h.59cgh5j3n0md) 103

[Two Main areas of inferential statistics](#_heading=h.spn7f8gbm2cj) 103

[Confidence Intervals of Mean, Proportion, and Variance](#_heading=h.j6sahogb1vw7) 106

[Mean:](#_heading=h.yd3onbxpnybi) 106

[Proportion:](#_heading=h.b99cd2sgrhn9) 108

[t-Distribution:](#_heading=h.lq6xl63qdnc2) 109

[Chi Square Distribution for Variance:](#_heading=h.h9v8a3c36i1s) 111

[Confidence Intervals Formulas](#_heading=h.mq1htd8cu5b0) 114

[Critical Region up Close](#_heading=h.grafeotblxor) 118

[One-tailed tests:](#_heading=h.lfljjlu1c8j4) 118

[Two-tailed tests:](#_heading=h.pl525og96iud) 119

[Types of Errors](#_heading=h.x9nnfm87udt1) 119

[Type I error:](#_heading=h.pnaq195vnaba) 120

[Type II error:](#_heading=h.54b9e6pqmgk0) 121

[Why are Type I and Type II Errors Important?](#_heading=h.rlbb1jwxunlc) 121

[Probabilities of Errors in our Example:](#_heading=h.l3t4zto27ucm) 122

[Power of Hypothesis Test:](#_heading=h.ev8y6u8zw2eb) 123

[Common Test Statistics for Inferential Techniques](#_heading=h.or2xmsudcky3) 124

[Types of t – tests](#_heading=h.4flv3trjt0ct) 124

[Two-Sample t-test for Paired Data:](#_heading=h.wp1p8uy5c5m4) 124

[Two-Sample t-Test for Unpaired Data:](#_heading=h.pjj6l2a1h69s) 127

[Confidence Intervals](#_heading=h.x2tlldvqw4hv) 131

[AIM](#_heading=h.b61pw5u5t0f2) 132

[Importing the necessary packages](#_heading=h.xpus6lc8pxf0) 134

[Calculating probabilities using normal and standard normal distributions](#_heading=h.m4sruomotkbx) 134

[Calculating the confidence intervals](#_heading=h.ltnlh147piu2) 143

[Performing Test of Hypothesis](#_heading=h.qtwui8mbr5md) 148

[Assessment](#_heading=h.4ov04yvcjmu7) 156

[Choose the appropriate option](#_heading=h.l2qb4mlwjct2) 156

[Fill in the spaces with appropriate answers](#_heading=h.vhgminvoejwe) 157

[Programming Assignment](#_heading=h.1mmjp4z0spj8) 157

[Solutions for Assessment](#_heading=h.evtxl9by0v3r) 158

[Choose the appropriate options](#_heading=h.oym0obdf444y) 158

[Fill in the spaces with appropriate answers](#_heading=h.rzvc0miialvz) 158

[Programming Assignment Solution](#_heading=h.k68dbndxtj92) 158

[**CHAPTER 4: ANOVA**](#_heading=h.vwmktrwlb2qd) **159**

[**Analysis of Variance**](#_heading=h.x0w8ry9edtbn) **159**

[Theory](#_heading=h.u57ak69yg9ee) 159

[Assumptions for ANOVA](#_heading=h.rfib7jgos4pw) 159

[Why ANOVA?](#_heading=h.7vv45j3ja4mv) 160

[Hypothesis test for 2 sample variances](#_heading=h.xt4n7tmm04b) 161

[Data](#_heading=h.tqrleqc9415f) 162

[What is a null and alternate hypothesis?](#_heading=h.p7e2ngq2oel9) 162

[What are numerator and denominator degrees of freedom?](#_heading=h.rrgx47xvxmq4) 162

[Reading an F-table](#_heading=h.qd53x7k0xzu0) 163

[Will you reject the null hypothesis or not?](#_heading=h.eqwcayox4d1m) 164

[What are the business implications?](#_heading=h.crw4fxln9v9) 164

[Comparing means](#_heading=h.7ahw0fyaq3e1) 164

[Comparing with population mean](#_heading=h.6ui3agpp687x) 164

[Why can’t we compute the t-test three times?](#_heading=h.wt1rnns477za) 165

[What changed?](#_heading=h.wz7igoj2gfkp) 166

[ANOVA: Analysis of Variance is variability ratio.](#_heading=h.n576n93o2k35) 167

[Solutions:](#_heading=h.c2sed0l3ps3u) 181

[Null and Alternative hypothesis](#_heading=h.tj6dh669cri1) 181

[Observed Frequencies](#_heading=h.tj9i3s25nbwe) 182

[Expected Frequencies](#_heading=h.69cm0xb2oao4) 182

[Chi-square](#_heading=h.laz4tapmduui) 182

[Case Study](#_heading=h.3vdjpqewznr0) 184

[Task 2: Two way ANOVA](#_heading=h.8lak4wcbuuo5) 191

[Task 3: Chi-square test of independence](#_heading=h.ctlu8qi4zu6k) 192

[Assessment](#_heading=h.4qlcgxln07b7) 194

[Choose the most appropriate option](#_heading=h.3os0llhyzm15) 194

[Fill in the spaces with appropriate answers](#_heading=h.hnvd5zmjvbtv) 195

[Programming Assignment](#_heading=h.ypaoe327k3n2) 195

[Solutions for Assessment](#_heading=h.cv2024jcexba) 196

[Choose the appropriate options](#_heading=h.p7r3npc0pdkg) 196

[Fill in the spaces with appropriate answers](#_heading=h.j9dl4lm3o7cp) 196

[Programming Assignment Solution](#_heading=h.d59djkifgpif) 196

[**CHAPTER 5: Data Cleaning and Exploratory Data Analysis**](#_heading=h.el54ufhbqwm0) **197**

[Theory](#_heading=h.ngvz7inggp82) 197

[Case Study](#_heading=h.aspoem61k4d4) 197

[Problem Definition](#_heading=h.15burvkdeb3) 197

[Imputing Missing Values](#_heading=h.ogk69wkz2oo4) 203

[Univariate analysis (Single variable plots)](#_heading=h.671mvg3wck24) 205

[Removing Outliers](#_heading=h.5qt4tnj08gb) 209

[Bivariate Analysis](#_heading=h.ch6yxg6ci5v5) 210

[Correlations between Features and Target](#_heading=h.i9zco8xp4bve) 213

[Multivariate Analysis](#_heading=h.otxbdi9yaka0) 217

[Assessment](#_heading=h.hffziwfwdj77) 224

[Choose the appropriate option](#_heading=h.2p113gwgrpc9) 224

[Programming Assignment](#_heading=h.t3gbyk26ujei) 225

[Solutions for Assessment](#_heading=h.957nifucmkve) 226

[Choose the appropriate options](#_heading=h.epehh2a4egah) 226

[Programming Assignment Solution](#_heading=h.wyln5r7ax4ml) 226

[**CHAPTER 6: Regression Analysis - Part I**](#_heading=h.90agk616jlca) **229**

[Theory](#_heading=h.3ngr6gldjbmv) 229

[The use of Regression Analysis](#_heading=h.k2y0axdw43rd) 229

[Linear Regression](#_heading=h.dwkx3kcdwbbo) 230

[Problem Statement](#_heading=h.g2sjxhnqyr16) 231

[Ordinary Least Squared Estimation](#_heading=h.ywd38pg4tt1a) 234

[The Mathematics involved](#_heading=h.xdms2js53ch5) 234

[Derivation of OLS by Minimizing Errors](#_heading=h.zddutj4hwv7p) 235

[Interpreting the model](#_heading=h.urq8g560r5qk) 236

[Prediction using the model](#_heading=h.1f2cj1b5ea67) 236

[Model Confidence](#_heading=h.2hvsdv52kbi7) 238

[Hypothesis Testing and p-values](#_heading=h.g0c6glicu8mh) 239

[Assumptions of OLS Regression](#_heading=h.7o2reruim5og) 242

[What if these assumptions are violated?](#_heading=h.d82hvdt51xp4) 243

[Interpretation of Regression Plots](#_heading=h.9t1ri98nlit9) 244

[Residual vs Fitted Values](#_heading=h.6m5as85qznxr) 245

[Normal Q-Q Plot](#_heading=h.e9ptkf64knw) 247

[Scale Location Plot](#_heading=h.goqzkga7snom) 247

[Residuals vs Leverage Plot](#_heading=h.tu9ieh1i8h8i) 248

[Model Evaluation Metrics](#_heading=h.23qlp9t9fk03) 249

[R Square/Adjusted R Square](#_heading=h.gnk1gx4v191u) 249

[Adjusted R Square](#_heading=h.px0s9migmuxv) 250

[Mean Square Error (MSE)](#_heading=h.wx8ay04giwgz) 251

[Root Mean Square Error (RMSE)](#_heading=h.jt6art4awg24) 252

[Mean Absolute Error](#_heading=h.wubkcdkni9lr) 252

[Overall Recommendation/Conclusion](#_heading=h.o3ia9qq5lkng) 253

[Practical Implementation](#_heading=h.aq2v33a7hiez) 253

[Import our libraries](#_heading=h.ar9l2n58v3k7) 253

[Load the Data into Pandas](#_heading=h.dgc0bcu3n2ti) 253

[Check for Multicollinearity](#_heading=h.isd6xxlsj6sh) 256

[Describe the Dataset](#_heading=h.4xp1lxa2e509) 259

[Filtering the Dataset](#_heading=h.1b8k6xcusltf) 260

[Build the model](#_heading=h.ag5x30oovu76) 260

[Exploring the output](#_heading=h.2hy01dvk0qcl) 261

[Evaluating the model](#_heading=h.hi0s12rjdsua) 262

[Using the Statsmodel](#_heading=h.z4n4rok5gaux) 262

[Checking for HeteroScedasticity](#_heading=h.rjcn73y7z1h) 263

[Checking for Autocorrelation](#_heading=h.cu4g8gcd6stq) 264

[Checking for Normally Distributed Residuals](#_heading=h.kvvx8cdt09gn) 266

[Checking the Mean of the Residuals Equals 0](#_heading=h.dnhlbzogmzhz) 266

[Measures of Error](#_heading=h.4h5rimd178dv) 267

[R-Squared](#_heading=h.ers182hngapj) 268

[Confidence Intervals](#_heading=h.gi2c348ghjq2) 268

[Hypothesis Testing](#_heading=h.sdp3ckvzq51) 269

[Create a Summary of the Model Output](#_heading=h.sh5adwty45mt) 270

[Remove the Insignificant Variables](#_heading=h.m7najmks77xx) 272

[Save the Model for Future Use](#_heading=h.1dde7vw6u210) 273

[Summary](#_heading=h.dgtwun7tirjl) 273

[Assessment](#_heading=h.45f7c8iu1r6b) 275

[Choose the appropriate option](#_heading=h.af9f58yeweal) 275

[Fill in the spaces with appropriate answers](#_heading=h.8f3y2mh0d74v) 276

[True or False](#_heading=h.37t8lnwq80m) 276

[Programming Assignment](#_heading=h.ca6ynyw10ecj) 276

[Solutions for Assessment](#_heading=h.ft0uvcw14ghu) 277

[Choose the appropriate options](#_heading=h.26ulotynbt8b) 277

[Fill in the spaces with appropriate answers](#_heading=h.l8vsr5m8ynj6) 277

[True or False](#_heading=h.xgy03k12ejv4) 277

[**CHAPTER 7: Regression Analysis – Part II**](#_heading=h.65yziot7xayx) **278**

[Theory](#_heading=h.o2p3ggpk83fr) 278

[Cost function Optimization Algorithm](#_heading=h.qe52eqe6yen4) 278

[What is Gradient Descent?](#_heading=h.quje536mafjh) 278

[Dummy variables versus quantitative explanatory variable](#_heading=h.xuq8z53xzjt1) 289

[Why do we need to perform feature scaling?](#_heading=h.wg00m685putq) 290

[Feature scaling techniques](#_heading=h.hdvy9frpqolb) 290

[Normalization](#_heading=h.da3d8m896rz7) 290

[Standardization](#_heading=h.625awf33w9bf) 291

[Backward Elimination](#_heading=h.uzkqrofagaxr) 292

[Forward Selection](#_heading=h.y9j38ewnczq) 293

[Stepwise Regression](#_heading=h.xaka3qdx523r) 293

[Underfitting and Overfitting](#_heading=h.3ngbcd1d51d9) 294

[Underfitting](#_heading=h.frdvxu791buo) 294

[Overfitting](#_heading=h.kyfzf7oha9wi) 294

[Bias and Variance Tradeoff](#_heading=h.md6ls9k4dxhz) 295

[Regularization](#_heading=h.2wmjyhffhnh2) 297

[Polynomial Regression](#_heading=h.2v7doovvgadn) 300

[When to use Polynomial Regression?](#_heading=h.ab0gdmp5kyty) 301

[AIM](#_heading=h.9px0slnv6wu7) 303

[Task 1): Handling Categorical Predictors with Two Categories](#_heading=h.8hf4d19xnda) 304

[Task 2) Handling Categorical Variables with More Than Two Categories](#_heading=h.3s4qd410g05a) 306

[Task 3) Regularization the algorithms](#_heading=h.ik4w6nnxjgrk) 308

[Summary](#_heading=h.o3s3pxra2vfv) 322

[Assessment](#_heading=h.j03a5h9oql2t) 323

[Choose the appropriate option](#_heading=h.id74mxobg9o4) 323

[Fill in the spaces with appropriate answers](#_heading=h.hdw1407ws2n4) 324

[True or False](#_heading=h.xtbw63qr13ws) 324

[Programming Assignment](#_heading=h.o8g65zicqpa1) 324

[Solutions for Assessment](#_heading=h.vlxdtyx0oyy4) 325

[Choose the appropriate options](#_heading=h.ouajadfgto01) 325

[Fill in the spaces with appropriate answers](#_heading=h.lpmdrx9p372j) 325

[True or False](#_heading=h.9c1vkjlxxscw) 325

# CHAPTER 1: INTRODUCTION TO STATISTICS

## Theory

This chapter provides a comprehensive introduction to Statistics, which is a building block for Data Analytics. You will understand the differences between population and sample, different types of variables as well as the types of descriptive statistics measures.



## Statistics – Big Picture

Statistics provides a method of organizing data to extract information on a wider and objective basis rather than relying on personal experience. It is a branch of mathematics working with:

* Data Gathering
* Data Understanding
* Data Analysis/Interpretation
* Data Presentation

## Basic Statistical Terminology



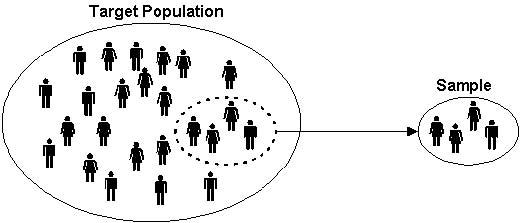


Image Source: <https://www.snapsurveys.com/blog/wpcontent/uploads/2011/08/target-population.jpg>.

A **population** is the collection of **all** items of interest to our study and is usually denoted with an uppercase **N**.

**Census** – It is a process of gathering data from the whole population of interest. For example, elections and 10-year census, *etc.*

A **sample** is a **subset** of the population and is denoted with a lowercase **n.**

**Survey** – It is a process of gathering data from the sample in order to make conclusions about a certain population. For example, opinion polls, and quality control checks in the manufacturing units, *etc*.

## Parameter and Statistic

**Parameter**: It is defined as a descriptive measure of the population. For example,

population mean - μ

population variance – σ2

population standard deviation - σ

**Statistic:** It is defined as a descriptive measure of the sample. For example,

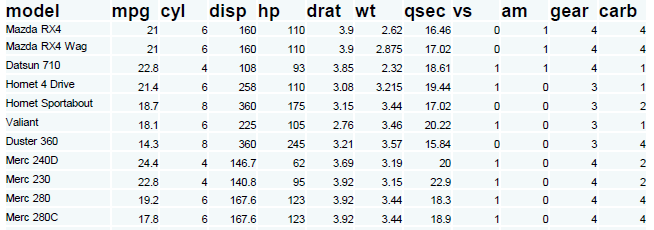
sample mean - xbar

sample variance - s2

sample standard deviation – s

## Variables and Data

**Example of data**



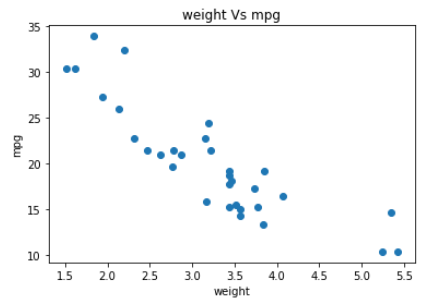
Source: MTCARS dataset. Data was extracted from the 1974 *Motor Trend* US magazine, and comprises fuel consumption and 10 aspects of the automobile design and performance for 32 automobiles (1973–74 models)

| **Marks  category** | **No. of students** | **% of students** | **Study hours per student** | **Mean marks** | **Standard deviation** |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
| **0 - 10** | 4 | 3.01 | 2.4 | 7.93 | 2.45 |
| **11 - 20** | 7 | 5.26 | 3.1 | 16.43 | 2.87 |
| **21 - 30** | 8 | 6.02 | 3.3 | 24.66 | 4.33 |
| **31 - 40** | 13 | 9.77 | 3.6 | 37.43 | 3.44 |
| **41 - 50** | 16 | 12.03 | 4.5 | 43.55 | 3.64 |
| **51 - 60** | 26 | 19.55 | 5.4 | 57.44 | 4.11 |
| **61 - 70** | 35 | 26.32 | 5.6 | 64.33 | 2.77 |
| **71 - 80** | 13 | 9.77 | 6.5 | 75.33 | 2.67 |
| **81 - 90** | 8 | 6.02 | 6.8 | 83.44 | 1.66 |
| **91 - 100** | 3 | 2.26 | 7.2 | 93.22 | 1.25 |

## Variables – Dependent and Independent

An independent variable (experimental or predictor) is a variable that can be manipulated in an experiment in order to observe its effect on the dependent variable(s) (Outcome).

In the above data example, the number of hours of study is the independent variable (you can manipulate it by increasing or decreasing your study duration)while the amount of marks obtained is an independent variable (you cannot directly control it, but you can influence it by manipulating the independent variable).



* Dependent variables on y – axis and Independent on x – axis
* Dependent variable is also called Target variable or Class variable.



## Data

Data is classified into two types, which are Categorical and Numerical data.

**Categorical Data** (Qualitative) represent types of data which may be divided into groups. For example, Car brands – Audi, BMW, and Mercedes. Gender – Male, Female.

Sometimes categorical data can hold numerical values (quantitative value), but those values do not have mathematical sense. For instance: birthdates, school postcodes. Here, the birthdate and school postcode hold the quantitative value, but do not give numerical meaning.

**Numerical Data** is also known as quantitative data and represents the numerical value (i.e, how much, how often, how many). Numerical data gives information about the quantities of a specific thing. It is further divided into two groups: Discrete and Continuous numerical data.

* **Discrete** data can be counted finitely. Those values cannot be meaningfully subdivided
  + Example: Number of students in a class: 10, 20, 30.
* **Continuous** data can be calculated, and it has an infinite number of probable values that can be selected within a given specific range.
  + Examples: Hight, Area, Distance, and Time, *etc.*

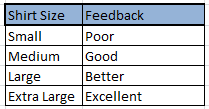
## Levels of Measurement Scales

In statistics, there are four data measurement scales: nominal, ordinal, interval, and ratio. These scales are the simple ways to sub-categorize different types of data.

* **Nominal scale:** The nominal scale could simply be called “labels”. These scales are mutually exclusive and none of them have any numerical significance. Here are some examples, below



* **Ordinal scale:** The order of the values is what’s important and significant, but the difference between each one is not characterized well. Here are some examples,



* **Interval Scale:** Interval scales are numeric scales in which we know both the order and the exact differences between the values. The classic example of an interval scale is Celsius temperature since the difference between each value is the same. For example, the difference between 60 and 50 degrees is a measurable 10 degrees, as is the difference between 80 and 70 degrees. The drawback of interval scales is that they don’t have a “true zero”. For example, there is no such thing as “no temperature”, on the celsius scale. In the case of interval scales, zero doesn’t mean the absence of value, but is actually another number used on the scale, like 0 degrees celsius. Negative numbers also have meaning. Without a true zero, it is impossible to compute ratios. With interval data, we can add and subtract, but cannot multiply or divide.
* **Ratio scale:** Ratio scales are the ultimate nirvana when it comes to data measurement scales because they tell us about the order and the exact value between units. They also have an absolute zero which allows for a wide range of both descriptive and inferential statistics to be applied. Ratio scales provide a wealth of possibilities when it comes to statistical analysis. These variables can be meaningfully added, subtracted, multiplied or divided (ratios).

## Descriptive Statistics



Descriptive statistics involves organizing, summarizing, and presenting data in an informative way. Descriptive statistics, unlike inferential statistics, seeks to ‘’describe’’ the data, without drawing interference or making comparisons with the data.

Example: You have exam results of a class of students. You can describe several attributes of this data set such as average score, median score, highest and lowest score, standard deviation, etc. Descriptive data provides you this type of information.

**Different types of Descriptive Statistics**

Descriptive statistics are broken down into two categories

* Measure of Central Tendency.
* Measure of Variability (Spread).



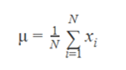
* **Mean:** Mean is a central tendency of the data i.e. a number around which a whole data set is spread out. In a way, it is a single number which can estimate the value of the whole data set. The mean has certain characteristics, such as
  + It is the most stable measure
  + It is affected by extreme values
  + It may not exist as data point in the set

Formula to calculate sample mean:



Here n is the size of the data set, x̄ is the sample mean, and x¡ the numbers in sequence, ∑ is the summation of the entire data set.

Similarly, for a population data of size N, the population mean can be calculated as:



Example: The systolic blood pressure of seven middle aged men is given as:

150, 123, 134, 170, 146, 124, and 113

The mean of these values is = (150+123+134+170+146+124+113)/7 = 137.14

Example: The systolic blood pressure of seven middle aged men with extreme value in:

150, 123, 134, 170, 146, 124, and 1113

The mean is = (150+123+134+170+146+124+1113)/7 = 280

We can see how mean is affected by one extreme value

* **Median:** Median is the value which divides the data into 2 equal halves based on the number of values i.e. number of terms on the right side of it is the same as the number of terms on the left side of it when data is arranged in either ascending or descending order. The median value has certain characteristics such as
  + It may not exist as a data point in the set..
  + It is influenced by the position of items, but not their values.
  + Median is not influenced by extreme values.

If the number of values in a data set is odd, then the median is the middle value. If the number of values in the data set is even, the median is the average of middle two values.

Example: if n is odd

The re-ordered systolic blood pressure data:

113, 124, 125, 132, 146, 151, and 170

The median here is 132

Example: if n is even

The re-ordered systolic blood pressure data:

113, 124, 125, 132, 146, 151, 161, and 170

The median here is (132+146)/2 = 139

* **Mode:** Mode is the most commonly occurring value. It has the following characteristics,
  + It exists as a data point.
  + It is useful for qualitative data.

Example: Six men with high blood cholesterol participated in the study to investigate the effects of diet on cholesterol levels. At the beginning of the study, their cholesterol levels (mg/dl) were as follows:

366, 327, 274, 292, 274, and 230

Rearrange the data in ascending order as follows:

230, 274, 274, 292, 327, and 366

The mode between the two men having the same cholesterol level = 274

But a dataset is possible without mode values if all the values appear the same number of times. If two values appeared at the same time and more than the rest of the values then the dataset is bi-modal. If three values appeared at the same time and more than the rest of the values, then the dataset is tri-modal and for n modes, that dataset can be termed as multimodal.

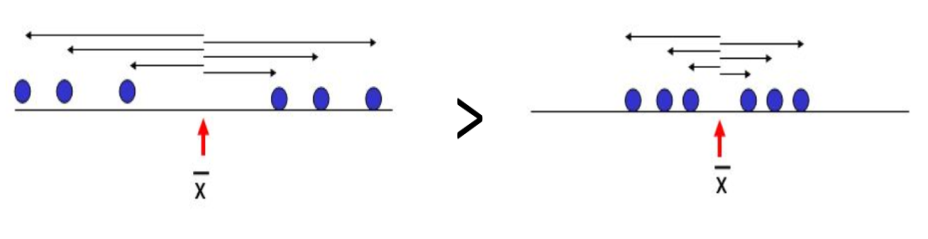


The mean, median, and mode are usually not by themselves sufficient measures to reveal the shape of a distribution of a data set. We also need a measure that can provide some information about the *variation* among data set values.

The measures that help us to know about the spread of a data set are called measures of dispersion.

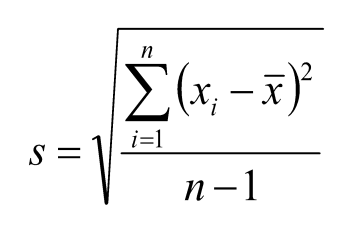
The measures of central tendency and dispersion taken together give a better picture of a data set.

* **Standard deviation:** Standard deviation is the measurement of the average distance between each quantity and the mean value. It shows us how data is spread out from mean. A low standard deviation indicates that the data points tend to be closer to the mean of the data set, while a high standard deviation indicates that the data points are spread out over a wider range of values.



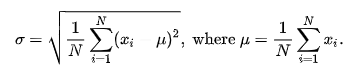
Here (left side) standard deviation is high because, from the mean (x̄), the points are distributed at a longer distance as compared to the right side, where the distance is a bit smaller.

Sample Standard Deviation is denoted by “S”



* + The sample standard deviation has the advantage of being in the same units as the original variables (x).
  + If the standard deviation is small, the data has low spread (*i.e.,* the majority of points fall very near the mean).
  + If standard deviation = 0, there is no spread. This only happens when all data items are of the same value.
  + The standard deviation is significantly affected by the outliers and skewed distributions.

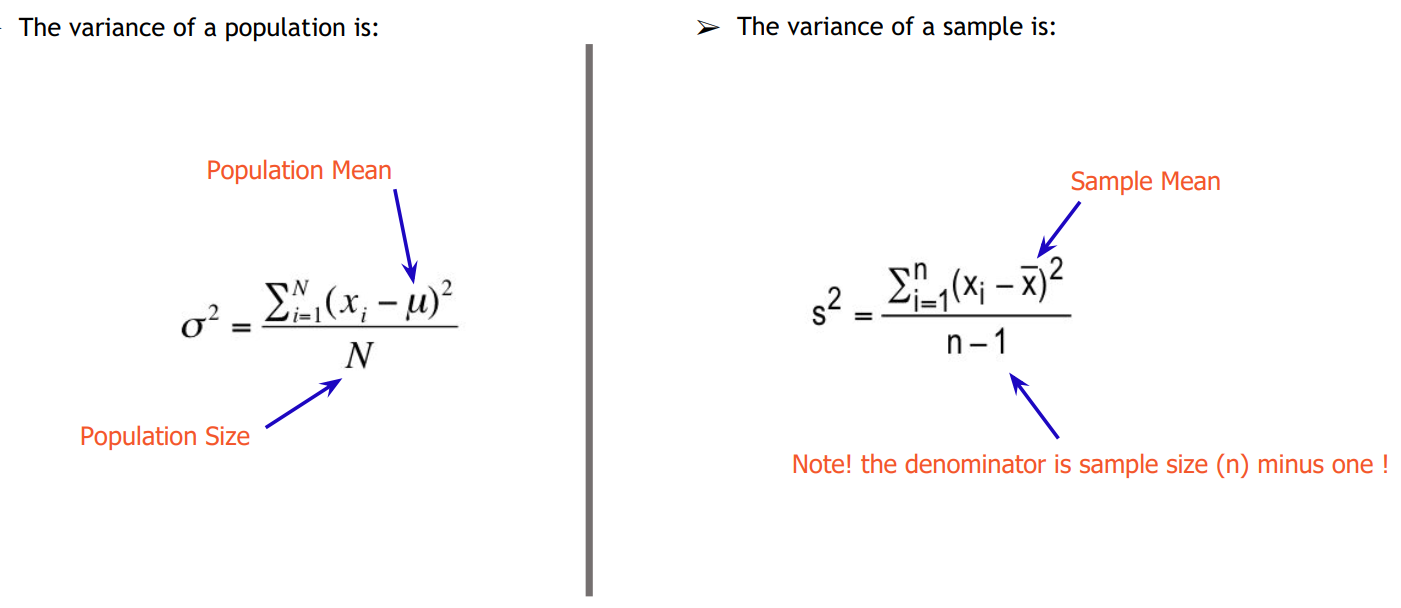
Population Standard Deviation is denoted by “σ” (sigma)



* **Variance:** Variance is the square of average distance between each quantity and mean. In simple words, it is the square of standard deviation.



The variance of Population and Sample



* **Range:** Range is one of the simplest techniques of descriptive statistics. It is the difference between the lowest and highest value.
  + It has the advantage of being easy to calculate.
  + It is implemented for both “best” or “worst” case scenarios.
  + It is very sensitive to extreme values.

Example: the minimum and maximum blood pressure are 113 and 170, respectively. Hence the range is 57.

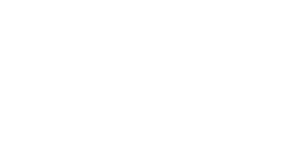
* **Percentile:** Percentile is a way to represent the position of a value in a data set. To calculate percentile, values in a data set should always be in an ascending order.

Example:

12, 24, 41, 51, 67, 67, 85, 99

The median 59 has 4 values, which are smaller than it in a dat set of 8 values. Thus, we can say that 59 is 50th percentile because 50% of the total terms are less than 59. In general, if **k** is **nth** percentile, it implies that **n%** of the total terms are less than **k**.

* **Quartile:** In statistics and probability, quartiles are values that divide your data into quarters, after the data is sorted in an ascending order.



* + There are three quartile values. First quartile value is at 25th percentile, second quartile is at 50th percentile, and third quartile is 75th percentile of the total.
  + Second quartile (Q2) is the median of the whole data.
  + First quartile (Q1) is the median of the upper half of the data.
  + Third quartile (Q3) is the median of the lower half of the data.
  + Quartiles exclude extreme values scientifically.

Example: Points scored per games are:

12, 24, 41, 51, 67, 67, 85, 99, 115

So here, by analogy,

Middle quartile (50th percentile, Q2) = 67 is the 50th percentile of the whole data.

First quartile (25th percentile, Q1) = 41 is 25th percentile and median of upper half of the data.

Third quartile (75th percentile, Q3) = 85 is 75th percentile and median of the lower half of the data.

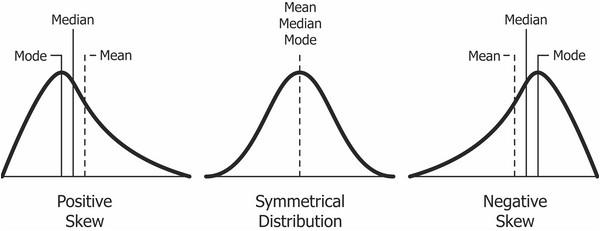
Interquartile range, IQR = Q3 – Q1 (Central 50% of data) = 85 – 41 = 44

* **Skewness:** Skewness is the measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. The skewness value can be positive or negative or undefined.

In a data set with perfectly normal distribution, the tails on either side of the curve are exact mirror images of each other, which is also called symmetrical distribution, where mean=median=mode.

When a distribution is skewed to the left, the tail on the curve’s left-hand side is longer than the tail on the right-hand side, and the mean is less than the mode and the median. This is also called the negative skewness.

When a distribution is skewed to the right, the tail on the curve’s right-hand side is longer than the tail on the left-hand side, and the mean is greater than the mode and the median. This is also called the positive skewness.

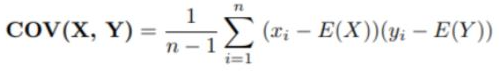




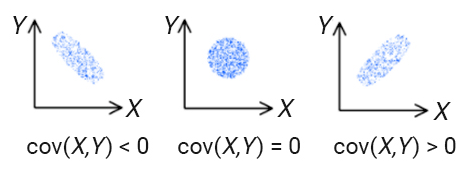
Covariance and correlation are two important concepts and are used in statistics and machine learning. One of the most commonly asked questions is the difference between these two terms and how to decide when to use them in a certain situation. Following are some definitions and mathematical formulas that will help you fully understand the concepts of covariance and correlation.

What is the Covariance Matrix?

A covariance matrix is used to study the direction of the linear relationship between different variables. Suppose we have two variables X and Y, and the covariance between these two variables is represented as Cov (X,Y). If ∑(X) and ∑(Y) are the expected values of the variables, then the formula for covariance can be represented as:



Here are some plots that highlight how the covariance between two variables looks like in different directions.



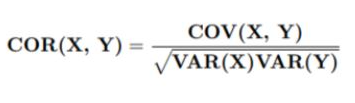
The covariance values of the variables can lie anywhere between -∞ and + ∞. A negative value indicates a negative relationship, whereas a positive value indicates a positive relationship between the variables.

When the unit of observation is changed for one or both of the two variables, the covariance value also changes. However, there is no change in the strength of the relationship.

To better understand the difference between covariance and correlation, let us understand what a correlation matrix is.

## What is a Correlation Matrix?

A correlation matrix is used to study the strength of a relationship between two variables. It shows the direction of the relationship in addition to showing the strength of it. The formula for the correlation can be represented as:



Where:

* + Var(X) = standard deviation of X.
  + Var(Y) = standard deviation of Y.

When the two variables move in the same direction, they are positively correlated. On the contrary, when the variables move in the opposite direction, they are negatively correlated with each other.

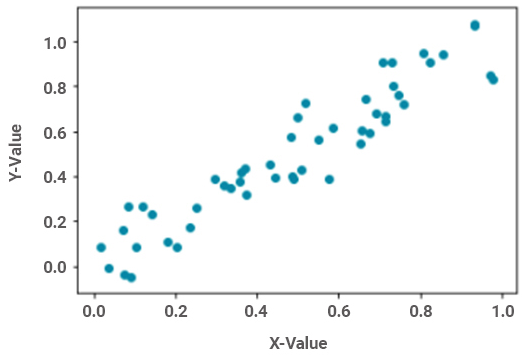


Fig: Positive relationship

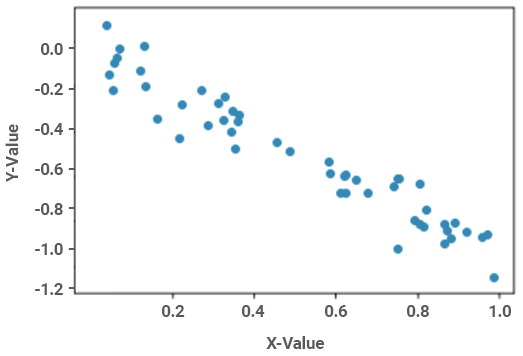


Fig: Negative relationship

The correlation value of two variables ranges from -1 to +1. A value close to +1 indicates a strong positive relation, whereas a value close to -1 indicates a strong negative correlation.

## AIM

The aim of the following section is to perform various exercises by writing python code so that we can practice hands-on descriptive statistics.

The labs for this chapter include the following exercises.

* Introduction to Python and Python versions
* Introduction to Anaconda Python Distribution and Installation
* What are Jupyter notebooks?
* How does a notebook work?
* Installing and launching the Jupyter Notebook.
* Importing necessary packages.
* Calculating the Central Tendencies.
  + Mean
  + Median
  + Mode
* Calculating Measure of dispersion
  + Standard Deviation
  + Variance
  + Range
  + Percentile
  + Quantiles
  + Boxplot
  + Skewness

We will be working with python3 and Jupyter notebook IDE in this chapter.

Note: This chapter assumes that you are familiar with python and its packages.

## Introduction to Python and Python versions

Python is an interpreter, object-oriented, high-level programming language with dynamic semantics. Its high-level built-in data structures, combined with dynamic typing and dynamic binding; makes it very attractive for Rapid Application Development. It is also effective for use as a scripting or glue language to connect existing components together. Python's simple, easy to learn syntax emphasizes readability and reduces the cost of program maintenance. Python supports modules and packages, which encourage program modularity and code reuse. The Python interpreter and the extensive standard library are available in source or binary form without charge for all major platforms and can be freely distributed. Python is a great general-purpose programming language on its own, but with the help of a few popular libraries (numpy, scipy, and matplotlib) it has become a powerful environment for scientific computing.

**Python Versions**

**Python 2**

Published in late 2000, Python 2 signaled a more transparent and inclusive language development process than the earlier versions of Python with the implementation of PEP (Python Enhancement Proposal), a technical specification that provides information to Python community members and describes a new feature of the language.

Additionally, Python 2 included many more programmatic features including a cycle-detecting garbage collector to automate memory management, increased Unicode support to standardize characters, and list comprehensions to create a list based on existing lists. Python 2 continued to develop with addition of more features including unifying Python’s types and classes into one hierarchy in Python version 2.2.

**Python 3**

Python 3 is regarded as the future of Python and is the version of the language that is currently in development. A major overhaul, Python 3 was released in late 2008 to address and amend the intrinsic design flaws of previous versions. The focus of Python 3 was to clean up the code base and remove redundancy, making it clear that there was only one way to perform a given task.

Since its development, Python 3 has undergone further modifications such as changing the print statement into a built-in function, improved ways of dividing the integers, and providing more Unicode support.

At first, Python 3 was slowly adopted due to the incompatibility of the language with Python 2, requiring people to make a decision as to which version of the language to use. Additionally, many package libraries were only available for Python 2. However, with time the development team behind Python 3 has reiterated that there is an end of life for Python 2 support, more libraries have been ported to Python 3. The increased adoption of Python 3 can be shown by the number of Python packages that now provide Python 3 support, which at the time of writing includes 339 of the 360 most popular Python packages.

**Python 2.7**

Following the 2008 release of Python 3.0, Python 2.7 was published on July 3, 2010 as the last of the 2.x releases. The intention behind the launch of Python 2.7 was to make it easier for Python 2.x users to port features over to Python 3 and provide additional compatibility between the two versions. This compatibility support included enhanced modules for version 2.7 like unit test to support test automation, argparse for parsing command-line options, and more convenient classes in collections.

Because of Python 2.7’s unique position as a version in between the earlier iterations of Python 2.0 and Python 3.0, it has persisted as a very popular choice for the programmers due to its compatibility with many robust libraries. When we talk about Python 2.0 today, we are typically referring to the Python 2.7 release, since that is the most frequently used version.

Python 2.7, however, is considered to be a legacy language and its continued development, which today mostly consists of bug fixes, will cease completely in 2020.

**Key Differences**

While Python 2.7 and Python 3 share many similar features, they should not be thought of as entirely interchangeable. Though you can write good codes and useful programs in either version, considerable differences exist in the code syntax and handling.

## Anaconda Python Distribution

Anaconda is an open-source package manager, environment manager, and a publicly-available distribution of the Python and R programming languages. It is commonly used for large-scale data processing, scientific computing, and predictive analytics. It is designed for serving data scientists, developers, business analysts, and those working in DevOps.

Anaconda offers a collection of over 720 open source packages and is available in both free and paid versions. The Anaconda distribution ships with the conda command-line utility. More can be learned about Anaconda and conda by reading the Anaconda Documentation pages.

## Why Anaconda?

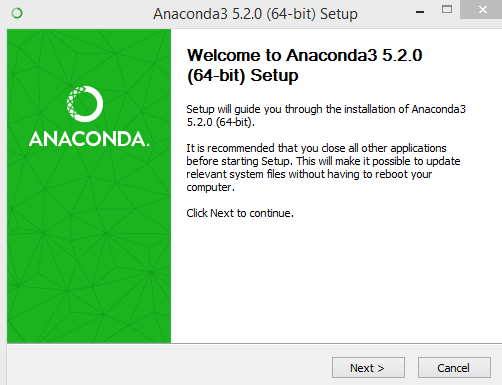
* It allows user level installation of the version of python.
* It enables the installation/update packages completely independent of system libraries or admin privileges.
* The Conda tool installs binary packages, rather than requiring compile resources like pip - again, handy if you have limited privileges for installing necessary libraries.
* It mostly eliminates the headaches of trying to figure out which version/release of package X is compatible with which version/release of package Y, both of which are required for the installation of package Z.
* It comes either in full-meal-deal version, with numpy, scipy, PyQt, spyder IDE, *etc*. or in minimal / alacarte version (miniconda) where you can install what you want, when you need it.
* It carries no risk of messing up required system libraries.

## Anaconda Installation

* [Download the Anaconda Installer(Windows X86\_64)](https://repo.anaconda.com/archive/Anaconda3-5.2.0-Windows-x86_64.exe)

(Other OS https://repo.anaconda.com/archive/)

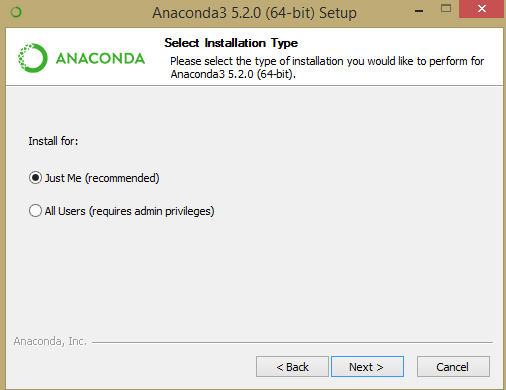
* **Optional**: Verify data integrity with MD5 or SHA-256 encryption algorithms. More info on hashes.
* Double click the installer to launch the Anaconda.



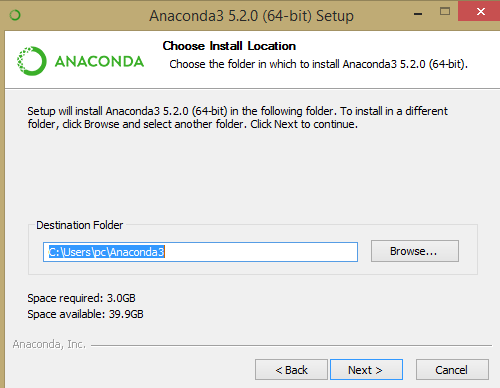
* Click Next



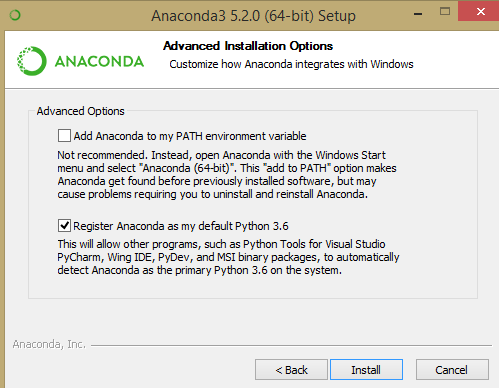
* Read the licensing terms and click I Agree.



* Select an installation option of “*Just Me*” unless you’re installing it for all the users (which may require the Windows Administrator privileges).



* Select a destination folder to install Anaconda and click next.



NOTE: Install Anaconda to a directory path that does not contain spaces or unicode characters.

NOTE: Do not install as ‘*Administrator’* unless the admin privileges are required.

* Choose whether to add Anaconda to your PATH environment variable. However, we recommend not adding Anaconda to the PATH environment variable, since this can interfere with other installed softwares. Instead, use Anaconda software by opening Anaconda Navigator or the Anaconda Command Prompt from the Start Menu.
* Choose the corresponding option to register Anaconda as your default Python 3.6. Unless you plan on installing and running multiple versions of Anaconda, or multiple versions of Python, you should accept the default option and leave this box checked.
* Click Install. You can click Show Details if you want to see all the packages Anaconda is installing.
* Click Next.
* After successful installation, you will see the “Thanks for installing Anaconda” message.
* You can leave the boxes checked “Learn more about Anaconda Cloud” and “Learn more about Anaconda Support” if you wish to read more about this cloud package management service and Anaconda support. Click Finish.
* After the installation is complete, verify it by opening the Anaconda Navigator, a program that is included with Anaconda. From your Windows Start menu, select the shortcut Anaconda Navigator. If the Navigator opens, it is an indication that you have successfully installed Anaconda.

## What are Jupyter notebooks?

* The notebook is a web application that allows you to combine the explanatory text, math equations, codes, and visualizations, all in one easily-shareable document.
* Recently, notebooks have quickly become an essential tool while working with data. You'll find their use in data cleaning and exploration, visualization, machine learning, and big data analysis. Typically you'd be doing this work in a terminal, either the normal Python shell or with IPython. Your visualizations would be in the separate windows, any documentation would be in separate documents, along with various scripts for functions and classes. However, with notebooks, all these features are in one place and can be easily explored and navigated together.
* Notebooks are also rendered automatically on GitHub. It’s a great feature that lets you easily share your work. There is also <http://nbviewer.jupyter.org/> that renders the notebooks from your GitHub repository or from notebooks stored elsewhere.

## How does a notebook work?

Jupyter notebooks grew out of the IPython project started by Fernando Perez. IPython is an interactive shell, like the normal Python shell but with additional features like syntax highlighting and code completion. Originally, notebooks worked by sending messages from the web app (the notebook you can only see in the browser) to an IPython kernel (an IPython application running in the background). The kernel would execute the code, and then sent it back to the notebook. The current architecture is similar, drawn out below.

The central point is the notebook server. You can connect to the server through your browser and the notebook is rendered as a web app. Your written code in the web app is sent through the server to the kernel. The kernel runs the code and sends it back to the server, and then any output is rendered back in the browser. When you save the notebook, it is written to the server as a JSON file with an .ipynb file extension.

The best part of this architecture is the fact that the kernel doesn't need to run Python. Since the notebook and the kernel are separate domains, code in any language can be transformed between them. For example, two of the earlier non-Python kernels were for the R and Julia languages. With an R kernel, code written in R will be sent to the R kernel where it is executed in a manner similar to the running of Python code on a Python kernel. IPython notebooks were renamed because notebooks became language agnostic. The new name Jupyter comes from the combination of Julia, Python, and R. If you're interested, here's a list of available kernels.

Another benefit is that the server can be run anywhere as it can be accessed via the internet. Typically, you'll be running the server on your own machine, where all your data and notebook files are stored. But you could also set up a server on a remote machine or cloud instance like Amazon's EC2. Then, you can access the notebooks in your browser from anywhere in the world.

## Installing and Launching the Jupyter notebook

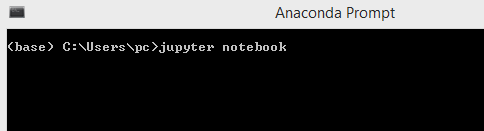
By far the easiest way to install Jupyter is with Anaconda. Jupyter notebooks are included in the distribution package. You'll be able to use notebooks from the default environment.

To install *Jupyter* notebooks in a conda environment, use **conda install jupyter notebook**.

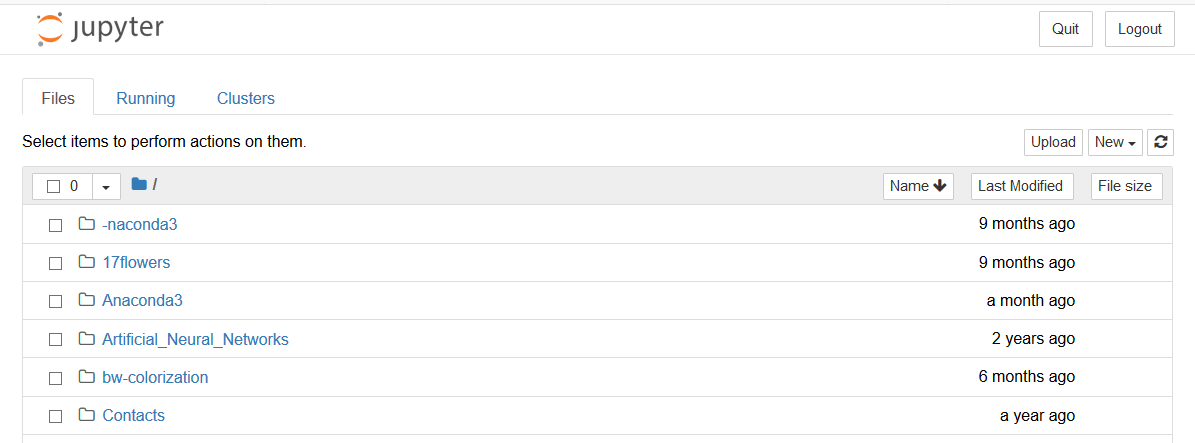
Moreover, the Jupyter notebooks are also available through pip with **pip install jupyter notebook**.

## Launching Jupyter notebook

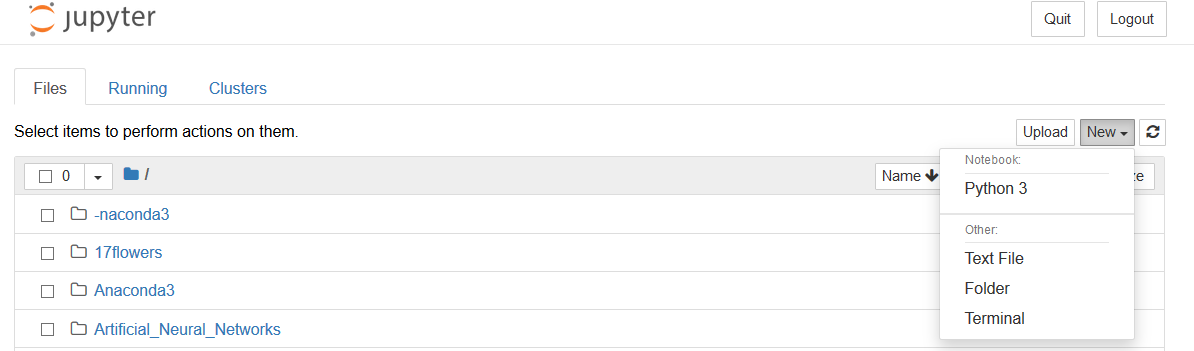
* To start notebook server, enter the following command in your Anaconda Prompt
  + $ jupyter notebook



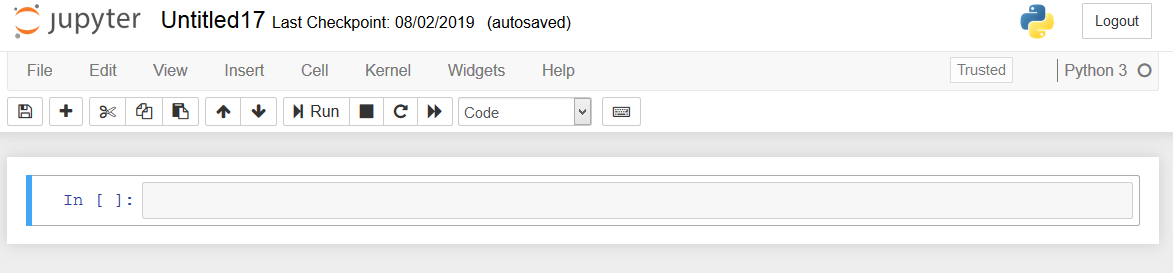
* When you run the command, the server home should open in your browser. By default, the notebook server runs at <http://localhost:8888>
* If you are not familiar with this, **localhost** means that your computer at port **8888** is communicating with the server. This communication carries on, as long as the server is still running, you can always come back to it by going to <http://localhost:8888> in your browser.
* If you start another server, it’ll try to use port **8888,** but since it is already occupied, the new server will run on port **8889.** Then, you’d connect to it at <http://localhost:8889>. Likewise, every additional notebook server will connect to a port with an incremented port number in this fashion.
* If you start your own server, it should look something like this:



* You can create a new notebook by clicking on *new and python3*.

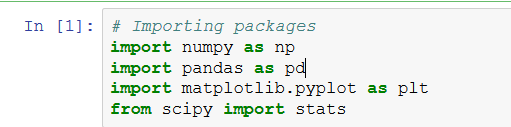


* Now we have launched our new notebook, we can go ahead with implementation.



## Importing the necessary packages

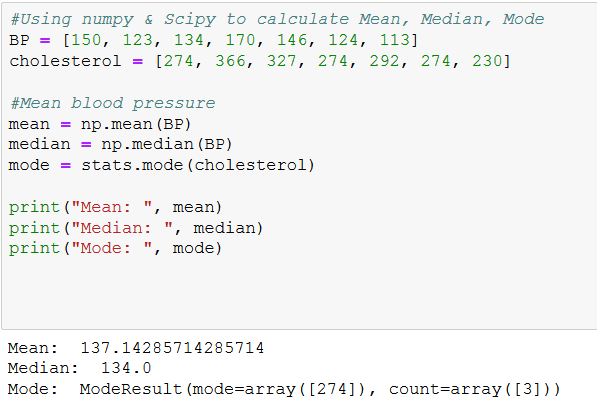
We always need to import the necessary packages required for the project or task.



## Calculating Central Tendencies

**Task 1: Mean, Median, and Mode**

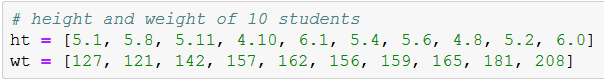
We will be using the data about “systolic blood pressure” and “blood cholesterol levels” to calculate central tendencies.



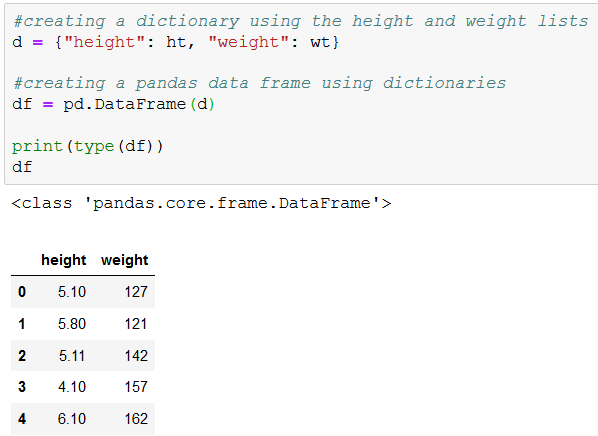
We first created BP (Blood Pressure) and plasma Cholesterol variables, then using numpy and scipy.stats packages we calculated the central tendencies (*i.e.* Mean, Median, Mode).

## Calculating Measure of Dispersion

We will be using a dataset which contains the height and weight of 10 students.

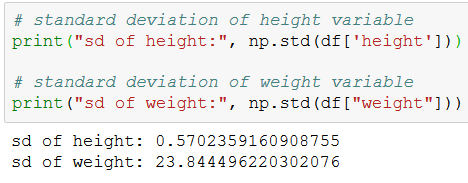


We have created a list of objects to store the data about height and weight.



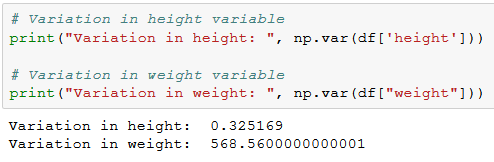
We have created Pandas DataFrame using the dictionary variable as input:

## Task 2: Standard Deviation



Standard deviation in the height variable is less than the standard deviation in weight variable. It means that the data points in the height variable are closer to the mean, when compared to the data points in the weight variable.

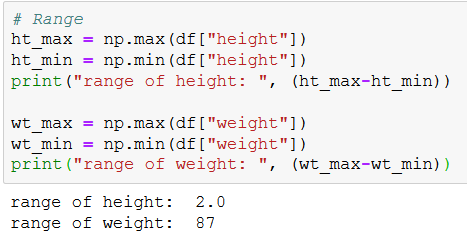
## Task 3: Variation



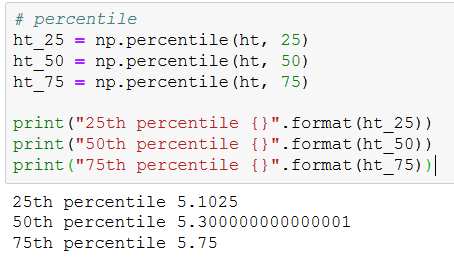
Variation in height variable is less compared to variation in weight variable.

## Task 4: Range and Percentile

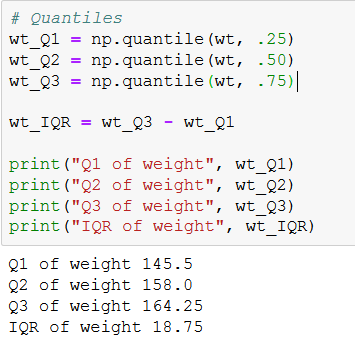
## Range



## Percentile



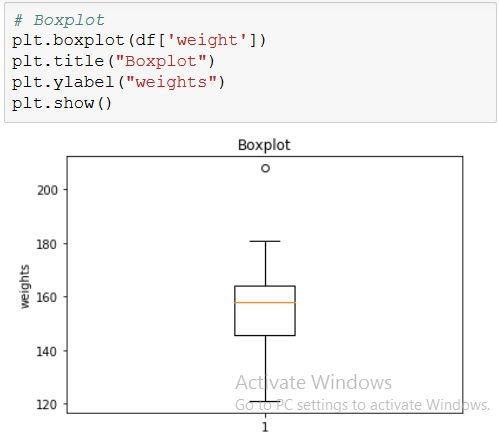
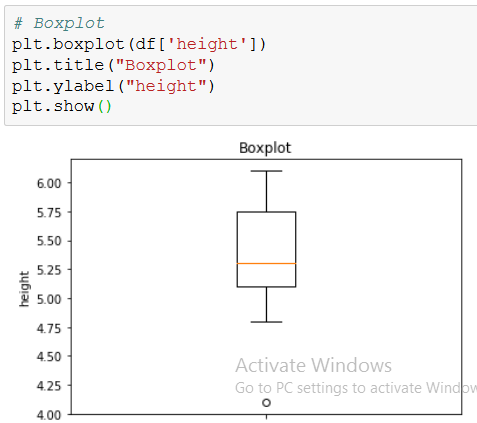
## Task 5: Quartiles



25% of the students have the body weight less than 145.5 pounds, 50% of the students have the body weight less than 158 pounds, 75% of students have the body weight less than 164.25 and finally, 50% of students weight lies in the range of 145.5 – 164.25 (i.e. Interquartile range = 18.75).

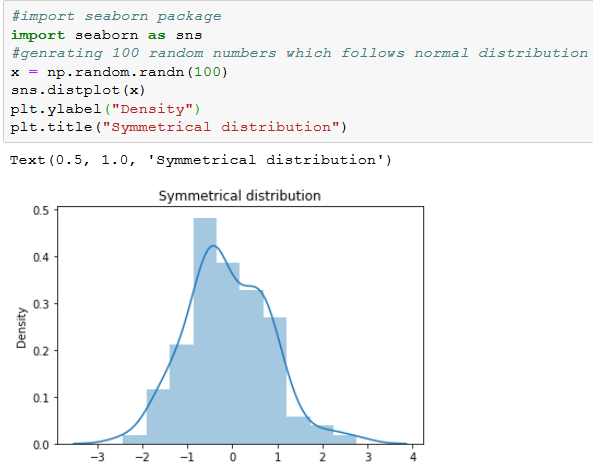
## Task 6: Boxplot

Boxplot is also called whisker plot. It is a robust way to identify extreme values (outliers). Boxplot contains upper whisker (*i.e.* Q3+IQR\*1.5) and lower whisker (*i.e.* Q1-1.5IQR). The data points which fall beyond those limits are considered as outliers.

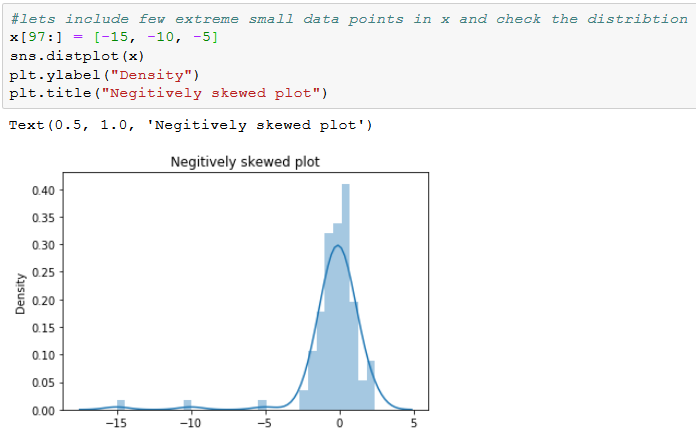
The above plots show an outlier in each of the height and weight variables. In the height variable, there is an extremely high value and in the weight variable, there is an extremely low value.

## Task 7: Skewness

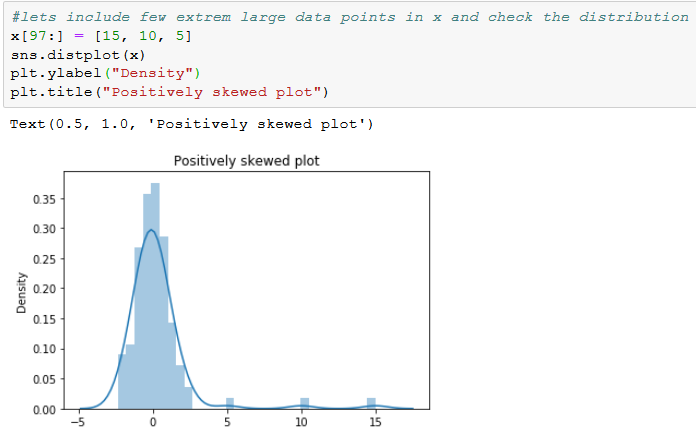


We have generated 100 random values using the np.random.randn (random normal distribution) function in the numpy package and have plotted a distribution plot. We will be further discussing normal distribution in future chapters.

## Negatively skewed distribution



## Positively skewed distribution



The distribution of the data will be affected due to a few extreme small (or) large data points called the *outliers*. We can see in the above examples that the symmetric distribution has been completely screwed with the inclusion of few outliers into the dataset.

## Case Study

We start our first chapter with a relatively easy problem about the reduction in the birth weights.

**Data description**

In 2015, 20.5 million newborns, an estimated 14.6 percent of all babies born globally that year suffered from the low-birth weight problem. These babies were more likely to die during their first month of life and those who survived faced lifelong complications of low birth weight; including a higher risk of stunted growth, lower IQ, and adult-onset chronic conditions, such as obesity and diabetes.

The dataset contains information about new born babies and their parents. The birth weights of the babies whose mothers smoked have been adjusted slightly to exaggerate the difference between mothers who smoked and the ones who did not smoke.

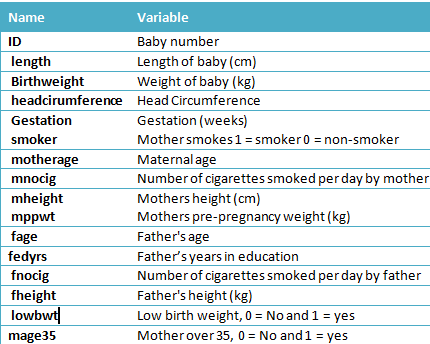
In this chapter, we will be exploring the data by using the descriptive statistics techniques, and then draw some meaningful insights from the data.

**Note**

The original dataset is available at the following link:

<https://www.sheffield.ac.uk/polopoly_fs/1.937185!/file/Birthweight_reduced_kg_R.csv>

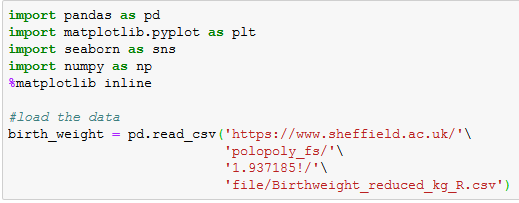
**general description of the variables**



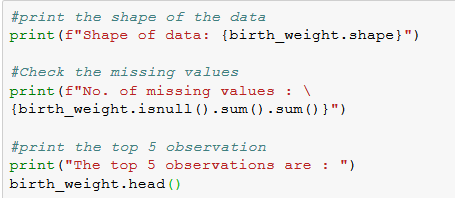
**Understanding the Data**

The main goal of the presented steps is to acquire the basic knowledge about the data, how its various features are distributed, and whether there are any missing volumes in it.

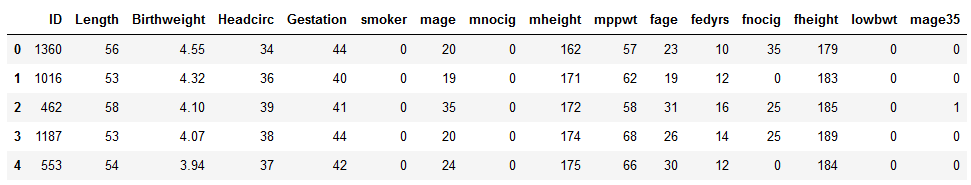
Please import relevant python libraries and data itself for the analysis.



A good starting practice is to check the size of the data we are loading, the number of missing values of each column, and explore the top 5 observations of the dataset:

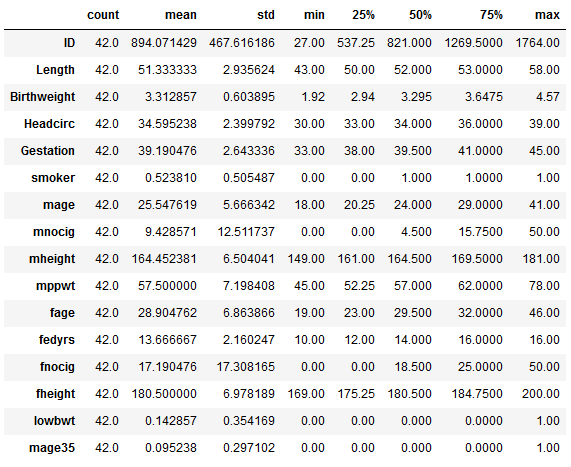






## Descriptive Statistics

In order to get simple statistics for the numerical columns, such as the mean, standard deviation, minimum and maximum values, and their percentiles, we can utilize the **describe** function on a **pandas dataset** object

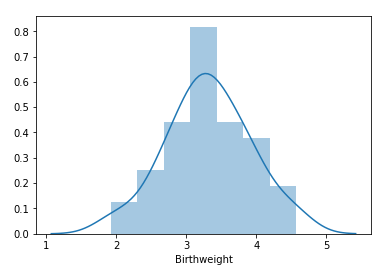


“*Birthweight*” is our dependent variable and the rest of the variables are independent. We want to analyze how the independent variables affect the *Birthweight* variable.

The mean birth weight is 3.312857 and the standard deviation is 0.603895. The maximum and minimum birth weights are 4.57 and 1.92, respectively.

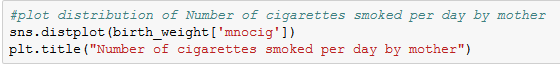
We can analyze the distribution of the birth weight variable.

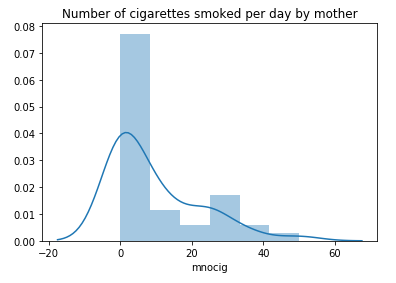




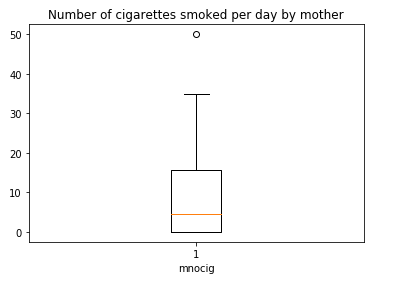
As you can see that “*Birthweight*” variable is symmetrically distributed around the mean.

Let us analyze the distribution of “**mnocig”** (Number of cigarettes smoked per day by mother) variable.





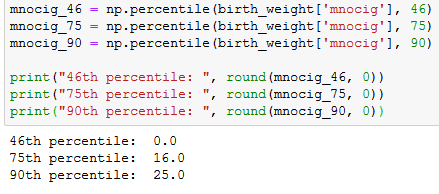
We can clearly observe that the data is positively or rightly skewed. Boxplot helps us to find out the presence of outliers in “*mnocig*” variable.



We can justify our assumption of outlier with the help of *boxplot*. One important insight we have to note here is that the mean of “*mnocig*” variable is 9.428571 and the median is 4.5.

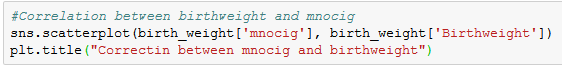
Mean is always influenced by an outlier but median is not. We can conclude that mean is not the best descriptive statistic in the presence of outliers in the data.

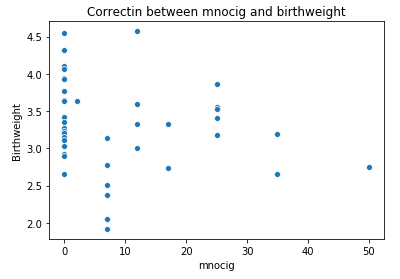
Thus, 50% of the mothers smoke less than 5 cigarettes per day, and the other 50% of the mothers smoke more than 5 cigarettes per day. To get more precise result, we can go with percentiles.

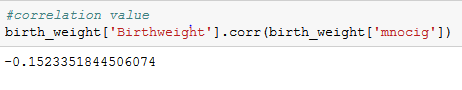


This means that 46% of mothers do not smoke cigarettes, 29% of mothers smoke less than or equal to 16 cigarettes per day, 15% of mothers smoke less than or equal to 25 cigarettes per day and finally, only 10% of mothers smoke more than 25 cigarettes per day.

Now it is time to understand the correlation between the dependent and independent variable.





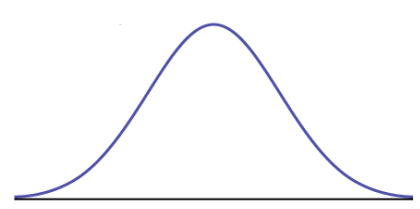


There is a very weak negative correlation between “*mnocig*” and “*Birthweight*”. We can understand that ‘mnocig’ does not play an important role in determining the “Birthweight” of the child based on the available data.

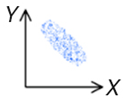
## Assessment

## Choose the appropriate option

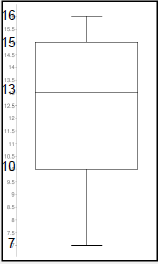
1. **Which of the following are the contents of descriptive statistics?**
2. Mean
3. Median
4. Percentile
5. Standard Deviation
6. All of the above
7. **Select an appropriate skewness of the curve From the below image.**



1. Positively Skewed
2. Negatively skewed
3. Symmetric
4. **From the below image select the appropriate relationship between the 2 variables:**



1. Negative Relationship
2. Positive Relationship
3. No Relationship
4. **The value of Correlation between two continuous variables lies between:**
5. 1 to 100
6. 0 to 1
7. -∞ to +∞
8. -1 to 1
9. **Which of the following statements are true?**



1. All of the students are less than 17 years old
2. Atleast75% of the students are 10 years old or older
3. There is only one 16 year old at the party
4. The youngest kid is 7 years old
5. Exactly half the kids are older than 13

## Fill in the spaces with appropriate answers

1. The method used to measure the spread of the data is \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.?
2. \_\_\_\_\_\_\_\_\_\_\_\_\_\_ is the robust visualization used to identify the outliers in the data.
3. Mean=Median=Mode in \_\_\_\_\_\_\_\_\_\_\_\_ distribution.
4. 4 different data measurement scales are\_\_\_\_\_\_\_\_\_\_\_, \_\_\_\_\_\_\_\_\_\_, \_\_\_\_\_\_\_\_\_\_\_, \_\_\_\_\_\_\_\_\_\_.
5. The 50th Percentile is also known as \_\_\_\_\_\_\_\_\_\_.

## Programming Assignment

Using the data in the below URL,

<https://www.sheffield.ac.uk/polopoly_fs/1.937185!/file/Birthweight_reduced_kg_R.csv>

By referring to the code used in the case study, perform the following tasks:

1. Find the distribution of ‘**fnocig’** (Number of cigarettes smoked per day by father) variable.
2. Find the correlation between ‘**Birthweight’** and ‘**fnocig’** variables.

**Solutions:** Refer to pages 44 - 45

## Solutions for Assessment

## Choose the appropriate options answers

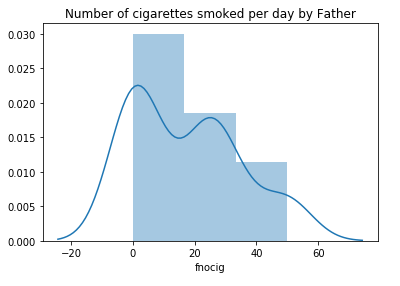
1. E
2. C
3. A
4. D
5. A)True B) False C) True D) True and E)True

## Fill in the spaces with appropriate answers:

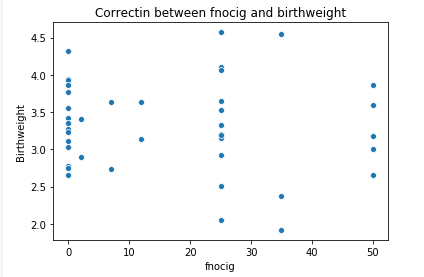
1. Measure of Variability (Spread / Dispersion)
2. Boxplot
3. Symmetric
4. Nominal, Ordinal, Interval, and Ratio
5. Median

## Programming Assignment Solution

Task 1)



Task 2)



# CHAPTER 2: PROBABILITY DISTRIBUTIONS

## Theory

Before we dive deeper into Inferential Statistics, let us first look into the Probability Distributions. This chapter covers most of the fundamental topics like Probabilities, Mutually Exclusive & Mutually Non-Exclusive Events, Conditional Probability, Random Variables, and Discrete Probability Distributions.



## Assigning Probabilities

Classical method – ***A prior or Theoretical***

Probability can be determined prior to conducting any experiment. It is formulated as follows:



**Experiment:** Tossing of a fair dice



**Outcome:** Possible Result of experiment {1, 2, 3, 4, 5, 6}

**Sample Space:** S = {1, 2, 3, 4, 5, 6}

**Event:** The thing of our interest, for example: getting a number ‘4’ on dice

P (4) = 1/6 = 1.6667

**Empirical Method –** ***A posteriori or Frequentist***

Probability can be determined post conducting a thought experiment. It is formulated as follows:

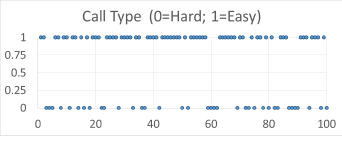


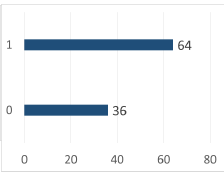
Example 1: Tossing of a weighted die…well! Or Even a fair die.

This is the most frequently used method in the statistical inference.

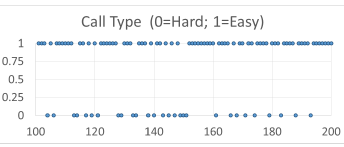
Example 2:

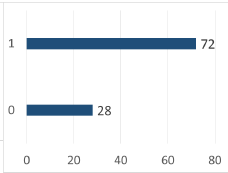
100 calls handled by an agent at a call center.



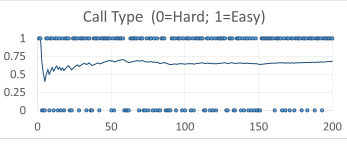


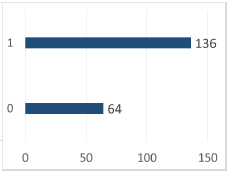
Next, 100 calls are handled by an agent at a call center.





Average over the long run





P (easy) = 0.7

**Subjective Method:** Based on feelings, insights, knowledge, etc. of a person.

What is the probability of India winning the upcoming series against England?

## Probability Terminology

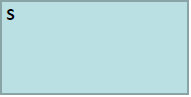
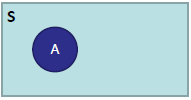
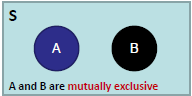
**Sample Space –** Set of all possible outcomes, denoted S.

Example: After two coin tosses, the set of all possible outcomes are {HH, HT, TH, TT}

**Event –** A subset of the samples space.

An Event of interest might be – HH.

## Probability – Rules

P(S) = 1 0 <= P (A) <= 1 P (A or B) = P (A) + P (B)

Area of the rectangle denotes sample space, and since probability is associated with area, it cannot be negative.



If A and B are mutually exclusive

P (A and B) = 0

But the probability of A or B is the sum of the individual probabilities.

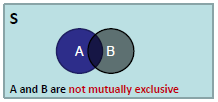
P (A or B) = P (A) + P (B)

Example: A card cannot be a King and Queen simultaneously, when it is picked from a deck of cards.

The Probability of a card being King is P (king) = 1/13

The Probability of a card being Queen is P (Queen) = 1/13

When we combine these two events: P (King or Queen) = (1/13) + (1/13) = 2/13



P (A or B) = P (A) + P (B) – P (A and B)



Events A and B cannot prevent the occurrence of one another, so we can say that events A and B have something common in them.

Example: In the case of rolling a dice, the probability of getting an “even number” and the probability of getting “less than 5” are not mutually exclusive and they are also known as a compatible event.

Let ‘A’ is denoted as event of getting an ‘even number’ and ‘B’ is denoted as event of getting ‘less than 5”.

The events of getting an even number (A) = {**2**, **4**, 6}

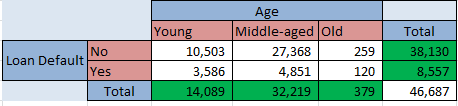
The events of getting less than 5 (B) = {1, **2**, 3, **4**}

Between the events A and B, the common outcomes are 2 and 4.

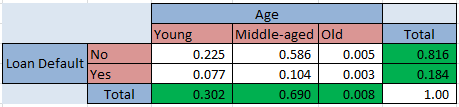
Therefore, events A and B are compatible or mutually non-exclusive events.

## Probability – Types

Contingency table summarizing two variables, Loan Default and Age:



Convert it into probabilities:



**Marginal Probability:** It is the probability describing a single attribute.

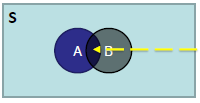


Example:

P (No) = 0.816

P (Old) = 0.008

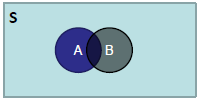
**Joint Probability:** It is the probability describing a combination of attributes.



Example:

P (Yes and Young) = 0.077

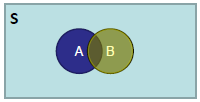
**Union Probability:** It is the probability describing a new set that contains all the elements that are in at least one of the two sets.



P (Yes or Young) = P (Yes) + P (Young) – P (Yes and Young) = 0.184 + 0.302 – 0.077 = 0.409

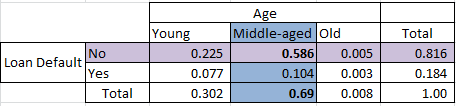


The sample space is restricted to a single row or column. This makes the rest of the sample irrelevant.



Example:

What is the probability that a person will not default on the loan payment **given** he/she is middle-aged?



P (No | Middle-Aged) = 0.586/0.690 = 0.85

Note that this is the rationale of Joint Probability to Marginal Probability, *i.e*.



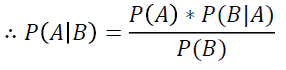
P (A/B) = P (A and B)/P (B) => P (A and B) = P (B) \* P (A/B)

Similarly

P (B/A) = P (A and B)/P (A) => P (A and B) = P (A) \* P (B/A)

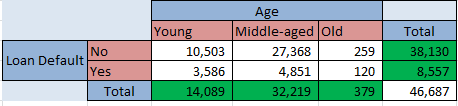
Equating, we get,

P (A/B) \* P (B) = P (A) \* P (B/A)



Now, given that the probability that someone defaults on a loan is 0.184, find the probability that an older person will default on the loan. Older people make up only 0.8% of the total clientele. P (Yes/Old) =?

P (Yes/Old) = (P (Yes) \* P (Old/Yes))/P (Old)



P (Yes) = 8557/46687 = 0.184

P (Old/Yes) = P (Old and Yes) / P (Yes) = 120/8557 = 0.014

P (Old) = 379/46687 = 0.008

P (Yes/Old) = (0.184 \* 0.014) / 0.008 = 0.32

The Probability that an older person defaulting on the loan is 32%

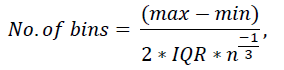
## Probability Distributions

**Histogram:** A series of contiguous rectangles that represent the frequency of data in the given class intervals.

How many class intervals?

Rule of thumb: 5-15 (not too many and not too few)

The Freedman-diaconis rule:



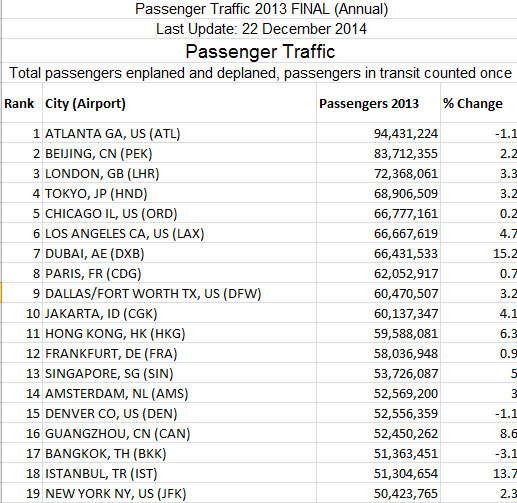
Where the denominator is the bin-width

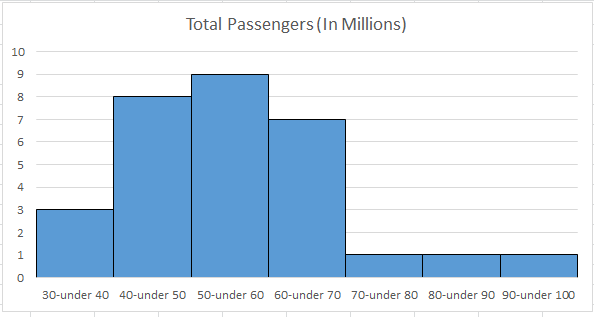
**Histogram – Excel**

Annual traffic data for the 30 busiest airports in the world – 2013

Source: <http://www.aci.aero/Data-Centre/Annual-Traffic-Data/Passengers/2013-final>

Last accessed: 22 December 2014







A random variable can take multiple values with different probabilities. The mathematical function describing these possible values along with their associated probabilities is termed as a *probability distribution*.



* The discrete random variable is a value that may take on only a countable number of distinct values.
* If a random variable can take only the finite number of distinct values, then it must be a *discrete random variable*.
* The probability distribution of the discrete variable is the list of the probabilities associated with every possible value.



Examples of discrete random variables:

* Number of children in the family.
* The Friday night attendance at a multiplex.
* The total number of defective light bulbs in the box

Suppose a random variable X may take k different values, with a probability that X=x, defined to be P(X = xi) = pi the probabilities pi must satisfy the following:

* 0 <= pi <= 1
* p1 + p2 + …. + pk = 1.

## Types of Discrete Probability Distributions

* Bernoulli Distribution.
* Binomial Distribution.
* Poisson Distribution.



A *Bernoulli* distribution is a discrete probability distribution of a random variable which takes 1 with probability ***P*** and the value 0 with probability q = 1-P.

**Notation: X ~ Ber(P)**

Example: The probability (p) of getting a head (“success”), while flipping a coin is 0.5. The probability of “*failure*” is 1-p (1 minus the probability of success, which also equals 0.5 for a coin toss).



A binomial distribution is the probability of a “success” or “failure” outcome in an experiment or survey that is repeated multiple times.

**Notation: X ~ Bio(n, P)**

**n: number of times the experiment runs**

**p: probability of one specific outcome**

**Probability Mass Function:**



Where:

* b = binomial probability.
* x = total number of “success”.
* P = probability of success on an individual trail.
* n = number of trails.

**Mean and Variance of Binomial distribution:**

**E (X) = np**

**Var (X) = npq**

**Criteria –** Binomial distribution must meet the following three criteria:

* The number of trials is fixed.
* Each trial is independent of others.
* The probability of “success” (trail, head, fail or pass) is the same from one trail to another.

Example 1: A coin is tossed 10 times. What is the probability of getting exactly 6 heads?

The number of trials (n) is 10.

The odds of success (“tossing a head”) is 0.5 (So 1-P = 0.5)

X = 6

P(X = 6) = 10C6 \* 0.5^6 \* 0.5^ (10-6)

= 210 \* 0.015625 \* 0.0625

P(X=6) = 0.2050

Therefore, the probability of getting exactly 6 heads is 20%

Example 2: Hospital records show that among the patients with a certain disease, ‘75%’ die of it. What is the probability that of six randomly selected patients, four will recover?

This is a binomial distribution because there are only two outcomes (die or survive).

Let X = number who recover.

Here n=6, X = 4

P=0.25 (success, i.e. they survive), q = 1-P = 0.75 (failure, i.e. die)

P(X=4) = 6C4 \* 0.25^4 \* 0.756-4

P(X=4) = 0.0329

The probability that out of six randomly selected patients, four will recover is 3.29%.



**Criteria –** Poisson distribution should meet the following criteria:

* The events can occur independently.
* An event can occur any number of times.
* The rate of occurrence is constant. i.e., the rate does not change based on time.

**Probability Mass Function:**



Where:

* The symbol “!” is a factorial.
* Μ (The expected number of occurrences) is sometimes written as λ. It is sometimes called the event rate or rate parameter.

**Mean and Variance of Poisson distribution:**

**E (X) =** 𝜆

**Var (X) =** 𝜆

Example 1: The average number of major earthquakes in your city is 2 per year. What is the probability that exactly 3 earthquakes will hit your city next year?

Step 1: Figure out the components you need to put into the equation.

* μ = 2 (average number of earthquakes per year, historically).
* x = 3 (the number of earthquakes we think might hit next year).
* e = 2.71828 (e is Euler’s number, a constant).

Step 2: Plug the values from step 1 into the Poisson distribution formula:

P(x, μ) = (2.71828^-2) \* (23) / 3!

= (0.13534) (8) / 6

P(x, μ) = 0.180

Hence, the probability of 3 earthquakes happening next year is 0.180, or 18%

Example 2: A life insurance salesman sells on average 3 life insurances per week. Use Poisson distribution to calculate the probability that,

1. In a given week he will sell how many policies.
2. In a given week, he will sell 2 or more policies but not more than five policies.
3. Assuming 5 working days per week, what is the probability that on a given day, he will sell one policy?

Here μ = 3

Solution 1: “Some policies” mean “1 or more policies”. We can work this out by computing to find 1 – “zero policies” probability:

P(X>0) = 1 – P(X)

P(X0) = e-3 30 / 0!

= 4.9787 \* 10-2

Therefore, the probability of 1 or more than 1 policy:

P(X>=0) = 1 – (4.9787 \* 10-2)

P(X>=0) = 0.95021

Solution 2: The probability of selling 2 or more policies but less than 5 policies are:

P (2 <= X < 5) = P(X2) + P(X3) + P(X4)



= 061611

Therefore, the probability of selling 2 or more policies but less than 5 policies is 61.16%.

Solution 3: The average number of policies sold per day is 3/5 = 0.6



= 0.32929

Hence, the probability of selling a policy on a given day is 0.32929 or 32.92%.

## Poisson distribution Vs Binomial distribution

In some cases, it can be challenging to find if you need Binomial distribution or Poisson distribution to fix your statistical problem. If you aren’t given a specific guideline from your instructor, use the following general guideline.

* If you are given an exact probability and you want to find the probability of an event happening for a certain number of times out of x number of times (i.e. 10 times out of 100 or 99 out of 1000), use the Binomial distribution.
* If your question has an average probability of an event per unit (i.e. per unit of time, cycle, event) and you want to find the probability of a certain number of events happening in a given period of time (or number of events), use the Poisson distribution.

## AIM

The aim of the following lab exercises is to perform the partial implementation of the Binomial and Poisson distributions by writing python code, so that we can get hands-on practice of Discrete Probability distributions

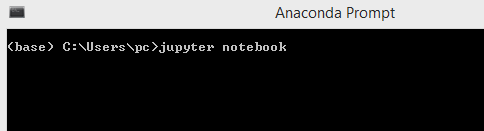
The labs for this chapter include the following exercises:

* Calculating probability using the Binomial distribution.
* Calculating probability using the Poisson distribution.

We will be working with python3 and the jupyter notebook IDE.

**Launching Jupyter notebook**

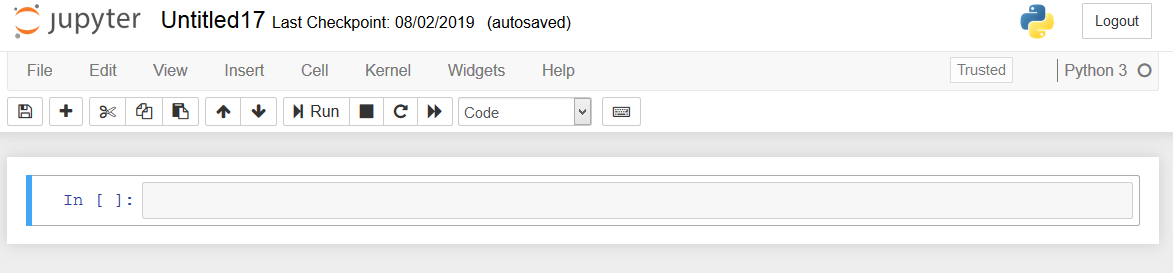
To start the notebook server, enter the following command in your Anaconda Prompt.



After starting the notebook server, create a new notebook by clicking on new and python3.

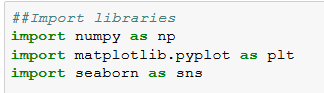


Now we have our new notebook, we can go ahead with the implementation.



## Importing the necessary packages

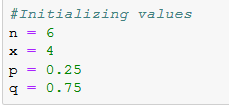
We always need to import the necessary packages required for the project or task.



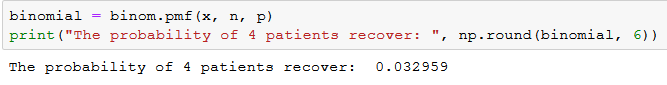
## Calculating probability using binomial distribution

**Task 1:** Hospital records show that among the patients with a certain disease, `75%` die of it. What is the probability that of six randomly selected Patients, four will recover from the disease?

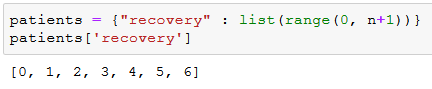
**Step 1**: We need to initiate the n, x, p, and q values.



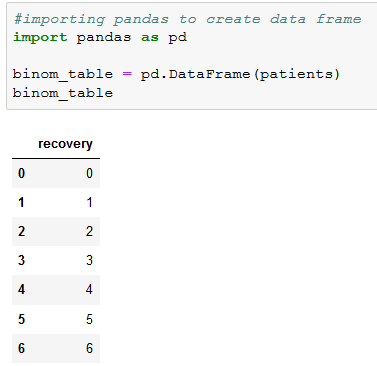
**Step 2:** Now, let’s use binom.pmf (binomial probability mass function) to calculate the probability that four out of six randomly selected patients will recover.



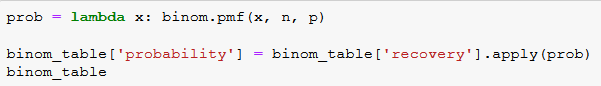
**Step 3:** Let’s see how the probabilities are distributed among the recovered patients.

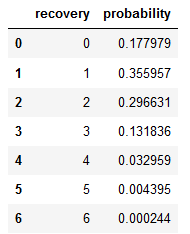


Created a dictionary to hold “*number of patients recovered*” values, we used ‘range’ function to generate the values between 0 and 6.



We created the data frame using pandas, DataFrame function.

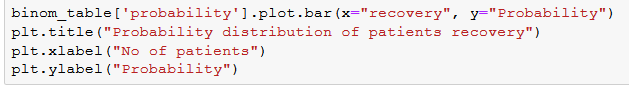


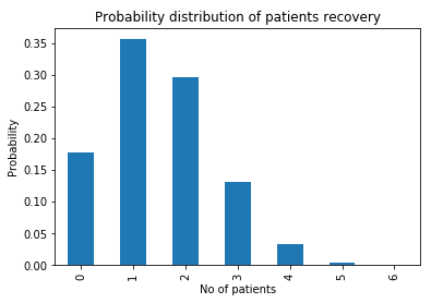


We defined a lambda function and named it as “prob”. “*A lambda function is a small anonymous function, which can take any number of arguments, but can only have one expression*”.

The objective of “prob” function is to calculate the probabilities, which use binomial probability mass function (binom.pmf).

We calculated the probability for each value of the “recovery” variable by passing the value to the “prob” function and saved the result to a new variable called “probability”.





**Interpretation:** The probability of recovery of four patients among six randomly selected patients is 3.29%. Similarly, the probability of recovery of one patient among six randomly selected patients is 35.59% and finally, the probability of recovery of all the six patients is almost 0%.

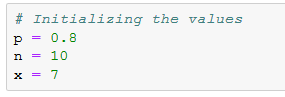
**Task 2:** In the old days, there was a probability of 0.8 to acquire success in any attempt to make a telephone call. Calculate the probability of having 7 successes (successful calls) in 10 attempts.

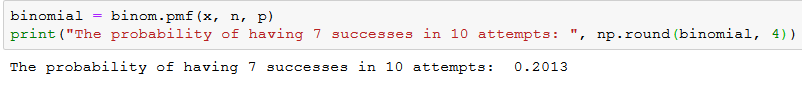
Solution:

Probability of success p=0.8, so q = 0.2

X = Success in getting through.

Probability of 7 successes in 10 attempts:





Therefore, there is a 20% probability of getting 7 successes in 10 attempts to make a call.

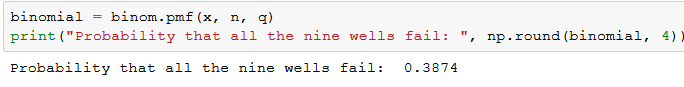
**Task 3:** A company drills 9 wildcat oil exploration wells, each with an estimated probability of success of 0.1. What is the probability that all nine wells fail?

Solution:

Probability of success p = 0.1, and failure is q = 0.9

n = 9





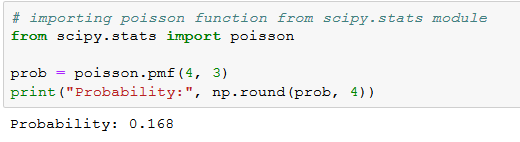
There is a 38.74% probability that all the nine exploratory drillings will fail.

## Calculating the probability using Poisson distribution

**Task 4:** John is recording birds in a national park, using a microphone placed in a tree. He is counting the number of times a bird is recorded singing and wants to model the number of birds singing in a minute. For this task, he’ll assume independence of the detected birds.

Looking at the data of the last few hours, John observes that on the average, 3 birds are detected in an interval of one minute. So, the value 3 could be a good candidate for the parameter of the distribution λ. His goal is to know the probability that a specific number of birds will sing in the next minute.

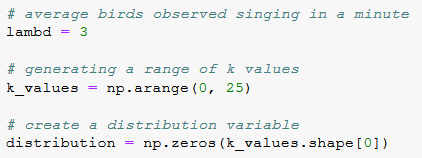
For instance, the probability of John observing 4 birds singing in the next minute would be,



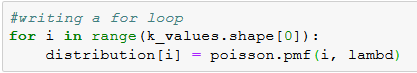
The probability of John observing 4 birds singing in the next minute is around 16.8%

Remember, the function poisson.pmf(k, lambda) takes the value of k and λ and returns the probability to observe k occurrences (*i.e.,* to record k birds singing).

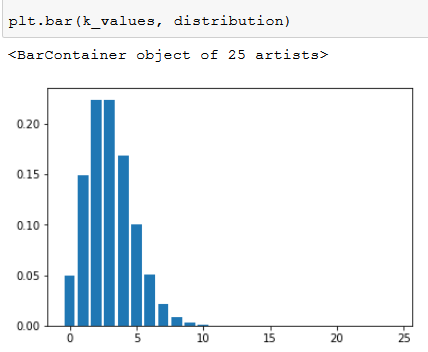
Let’s plot the distribution for the various values of k:



We created a “lambd” variable to hold average birds recorded in a minute, “k\_values” stores various values of k. Finally, we created distribution variable to store probability values.



We have passed the various k values and a lambda value to poisson.pmf though loop and saved the probabilities into the distribution variable.



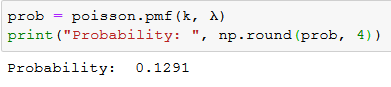
The probabilities corresponding to the values of k are summarized in the probability mass function shown in the above figure.

**Interpretation:** As shown, it is most likely that John will hear two to three birds singing in the next minute.

**Task 5:** If electrical power failures occur according to the Poisson distribution with an average of ‘3’ failures every twenty weeks, calculate the probability of occurrence of 1 failure during a particular week.



In the problem statement, λ value was given for 20 weeks. We calculated the average λ value per week as follows:



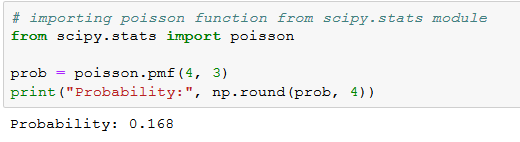
Therefore, the probability of 1 electrical power failure occurring in a week is 12.91%.

## Case Study

John is recording birds’ sounds in a national park, using a microphone placed in a tree. He is counting the number of times a bird is recorded singing, and wants to model the number of birds singing per minute. For this task, he will assume the independence of the detected birds.

Looking at the data of the last few hours, John observes that on average, 3 birds are detected making sounds in an interval of one minute. Thus, the value of 3 could be a good candidate for the parameter of the distribution λ. His goal is to determine the probability that a specific number of birds will sing in the next minute.

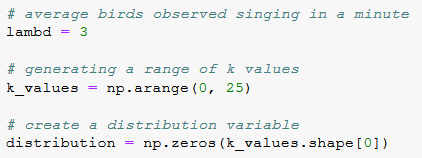
For instance, the probability of John observing 4 birds singing in the next minute would be:



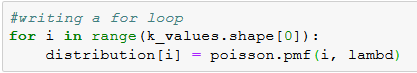
The probability of John observing 4 birds singing in the next minute is around 16.8%.

Remember that the function *poisson.pmf* (k, lambda) takes the value of k and λ and returns the probability to observe k occurrences (*i.e.*, to record k birds singing).

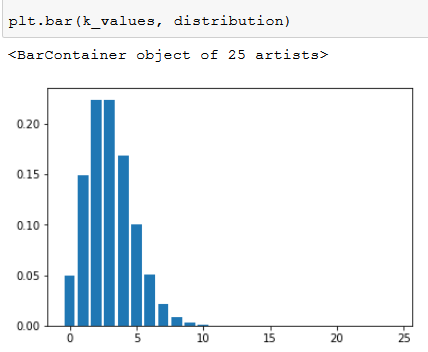
Let’s plot the distribution for various values of k:



We created a “*lambd*” variable to hold the average birds’ sounds recorded in a minute, “*k\_values*” stores various values of k and finally, we created the distribution variable to store probability values.



We have passed various k values and a lambda value to *poisson.pmf* through a loop and saved the probabilities into the distribution variables.



The probabilities corresponding to the values of k are summarized in the probability mass function shown in the figure.

**Interpretation:** You can see that it is most probable that John will record two or three birds singing in the next minute.

## Assessment

## Choose the appropriate option

1. Which of the following statements best describes the expected value of a discrete random variable?
   1. It is the geometric average of all possible outcomes.
   2. It is the weighted average over all possible outcomes.
   3. It is the simple average of all possible outcomes.
   4. None of the above.
2. Which of the following is not a true statement about the binomial probability distribution?
   1. Each outcome is independent of each other.
   2. Each outcome can be classified as either success or failure.
   3. The probability of success must be constant from trail to trail.
   4. The random variable of interest is continuous.
3. If n=10 and p = 0.8, then the mean of the binomial distribution is?
   1. 0.08
   2. 1.26
   3. 1.60
   4. 8.00
4. If the outcomes of a discrete random variable follow a Poisson distribution, then their
   1. Mean equals the variance.
   2. Mean equals the standard deviation.
   3. Median equals the variance.
   4. Median equals the standard deviation.
5. The sum of the product of each value of a discrete random variable X times its probability is referred to as its
   1. Expected value.
   2. Variance.
   3. Mean.
   4. Both (a) and (c)

## Fill in the spaces with appropriate answers

1. The \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ distribution can be used to approximate the binomial distribution when the number of trails is large and the probability of success is small (np <=7).
2. If n = 10 and p = 0.8, then the standard deviation of the binomial distribution is \_\_\_\_\_\_\_\_\_\_\_\_\_.
3. If two events A and B are mutually exclusive then P (A∩B) = \_\_\_\_\_\_\_\_\_\_\_\_.
4. A coin is tossed up 4 times. The probability that tails turn up in 3 cases is \_\_\_\_\_\_\_\_\_\_.
5. Mutually Exclusive events \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

## Programming Assignment

**Assignment 1**

Case study

At a Biscuit factory in Slough with 120 production workers, there is a 10% chance that a worker is absent on a given day. The probability that one worker is absent, assumes that it will not affect the probability that another worker is absent. The factory can operate on any given day as long as no more than 50 workers are absent on that day. What is the probability that any 2 out of 9 randomly chosen workers will be absent on the next Monday?

Solution:

This situation can be described by a Binomial Distribution since we have:

* A fixed number (9) of trails (workers).
* Each trial has two possible outcomes, “success” (the worker is absent) or “failure” (the work is not absent).
* The probability of success (0.1) is constant.
* The outcome of each trial is independent of the outcome of all the other trials.

Using the given values, calculate the probability that any 2 out of 9 randomly-chosen workers will be absent P (X =2)?

**Assignment 2**

**Case study**

The average daily sales volume of 60-inch 4K HD TVs at XYZ Electronics is five. Calculate the probability of XYZ Electronics selling nine TV sets today.

**Solution**

We have

* λ = 5, since five 60-inch TVs in the daily sales average.
* X = 9, we want to solve for the probability of nine TVs being sold.

Using Poisson distribution find the probability, P (X = 9).

**Solutions:** Refer to page 31

## Solutions for Assessment

## Choose the appropriate options answers

1. A
2. A
3. D
4. A
5. C

## Fill in the spaces with appropriate answers

1. Poisson
2. 1.26
3. 0
4. ½
5. Does not contain any common sample point

## Programming Assignment Solutions

Task 1)



Task 2)



# CHAPTER 3 – INFERENTIAL STATISTICS

## Theory

In the previous chapter, we studied the probability distributions and discrete random variables. Let us now deep dive and learn about continuous random variables, point estimate, confidence interval estimate, test of hypothesis, and t-test.

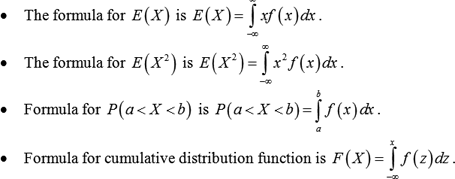
## Continuous Random Variable

A continuous random variable is a random variable, in which the data can take infinitely many values. Continuous random variables usually are the measurements.

Examples:

* The time required by the computer to process a certain program.
* The amount of rainfall in a certain city at a given time.
* The heat gained by a ceiling fan when it has worked for an hour.
* The amount of water passing through a pipe connected with a high-level reservoir.
* The time in which poultry will gain 1.5kg weight.

The probability function of the continuous random variable is ***probability density function*** (PDF) and represented by the area under a curve. It is formulated as follows:



## Types of Continuous Random Variables

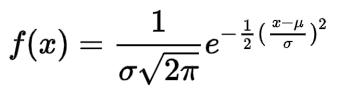
* Normal Distribution.
* Standard Normal Distribution.
* Uniform Distribution / Rectangular Distribution.



**Assumptions of Normal Distribution**

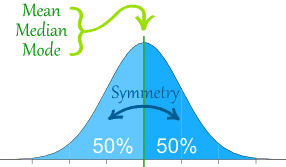
* It has a “Bell Shaped Curve”.
* Mean = Median = Mode.
* Zero Skew.
* Area under the curve is 1 (or) 100%.
* Notation: X ~ N (μ, σ2).

**Probability Density Function:**



Where:

* f(x) = probability density function.
* σ = standard deviation.
* μ = mean.



Source: [Image](https://www.google.com/url?sa=i&url=https%3A%2F%2Fdietassessmentprimer.cancer.gov%2Flearn%2Fdistribution.html&psig=AOvVaw1XgzS0U_iBQW9AG8E-q33r&ust=1612116874787000&source=images&cd=vfe&ved=2ahUKEwjms8S-ocTuAhWY_DgGHR_CARgQjRx6BAgAEAc)

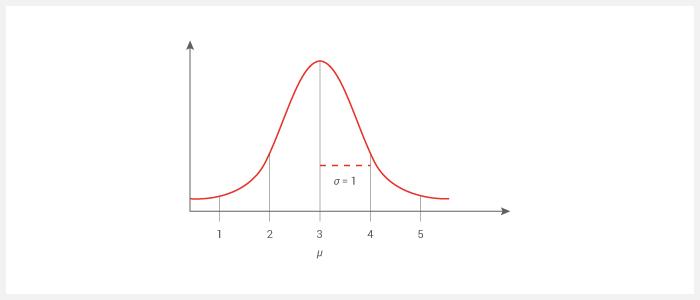
In the above image, we can view that the highest point is located at the mean μ, and the spread of the graph can be observed by the standard deviation σ.

Let us understand Probability Density Function with the easiest example where we can have the random variable X with distribution:

X = {1, 2, 3, 4, 5}

From the above data set, we get the mean (μ) = 3 and standard deviation (σ) = 1.

When we plot it, we get a few distributions like the ones shown in the below figure:



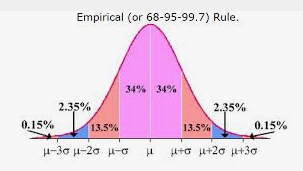
This Bell curve specifies the ***Gaussian / Normal distribution***.

**Empirical Formula**

The empirical rule states that for the Normal Distribution, nearly all observed data sets will fall within 3 standard deviations. The empirical rule is also known as **Three-Sigma** rule or **68-95-99.7** rule.

More specifically, the empirical rule predicts that:

* 68% of observations falls within the first standard deviation (µ ± σ).
* 95% of observations falls within two standard deviations (µ ± 2σ).
* 99.7% of observations falls within three standard deviations (µ ± 3σ).

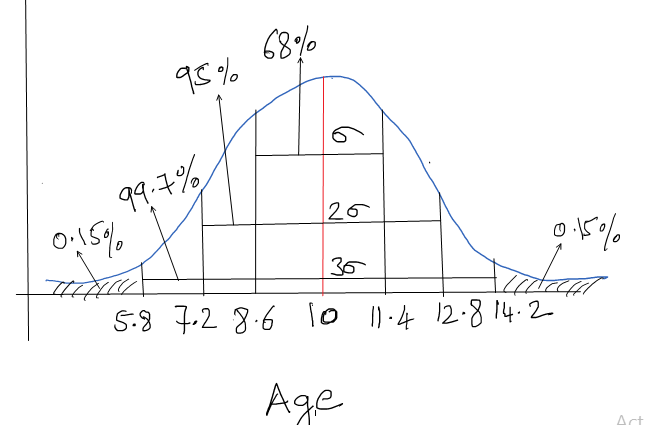


Source: <https://lh3.googleusercontent.com/gF03vD4TPhPzzIdJH1X701mkxrLVchWCU_6ZrSPLADfsFU-GRLDFNn7SR0QSZztPvyF8oA=s170>

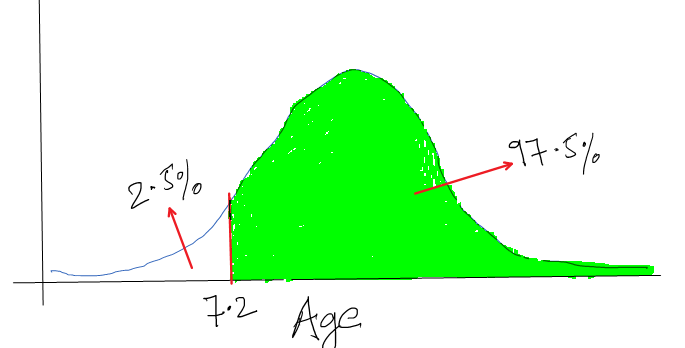
Example: The animal in the zoo has an average life expectancy of 10 years with a standard deviation of 1.4 years. Assume that the zookeeper attempts to figure out the probability of an animal living for more than 7.2 years.

The distribution looks as follows:

* One standard deviation (µ ± σ) (10 ± 1.4) = 8.6 to 11.4 years.
* Two standard deviations (µ ± 2σ) (10 ± 2\*1.4) = 7.2 to 12.8 years.
* Three standard deviations (µ ± 3σ) (10 ± 3\*1.4) = 5.8 to 14.2 years.



The empirical rule states that 95% of the distribution lies within two standard deviations. Thus 5% of the observations lies outside the two standard deviations, which include half the values above 12.8 years and the other half below 7.2 years.



95% + (5% / 2) = 97.5%

Thus, the probability of living more than 7.2 years is 97.5%.

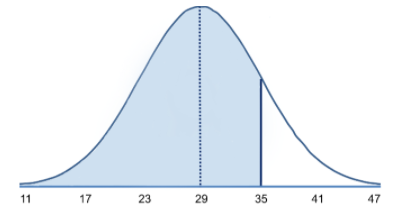
Example 2: Consider the weight values of a population of 15 years old males where the weight is normally distributed and has a mean value =29 and a standard deviation = 6.



The mean (μ = 29) is in the center of the distribution, while the horizontal axis is scaled in increments of the standard deviation (σ = 6). The distribution essentially ranges from μ - 3σ to μ + 3σ. It is possible to have weight values below 11 or above 47, but such extreme values occur very infrequently. To compute probabilities from the normal distributions, we will compute the areas under the curve. For any probability distribution, the total area under the curve is 1.

In the case of normal distribution, we know that the mean is equal to the median, so half (50%) of the area under the curve is above the mean and half is below, P (weight < 29) = 0.50. Consequently, if we select a kid at random from this population and ask his weight, what is the probability that his weight is less than 29kg? The answer is 0.50 or 50%, since 50% of the area under the curve is below the given weight value, which is 29. Note that with normal distribution, the probability of having any exact value is 0 because there is no area at an exact weight value, so in this case, the probability that his weight = 29 is 0, but the probability that his weight < 29 or the probability that his weight is < 29 is 50%.

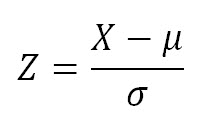
What is the probability that a 15-year-old male kid weighs less than 35? The probability is displayed graphically and represented by the area under the curve to the left of the value 35 in the figure below.



Note that weight = 35 is 1 standard deviation above the mean. In case of normal distribution approximately 68% of the area under the curve lies between (µ ± σ). Therefore, 68% of the area under the curve lies between 23 and 35. Moreover, we also know that the normal distribution is symmetric about mean, therefore P (29 < X < 35) = P (23 < X < 29) = 0.34. Consequently, P(X < 35) = 0.5 + 0.34 = 0.84. In other words, 68% of the area is between 23 and 35, so 34% of the area is between 29 and 35, and 50% is below 29. If the total area under the curve is 1, then the area below 35 = 0.50 + 0.34 => 0.84 or 84%.

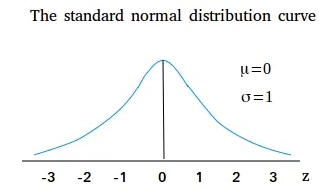


The normal random variable of a standard normal distribution is called a **Standard Score** or a **Z-Score**. Every normal random variable **X** can be transformed into a **Z-Score** using the following equation:



Where:

* X is a normal random variable.
* μ is the mean of X.
* σ is the standard deviation of X.



For the standard normal distribution,

* 68% of the observations lie within one standard deviation of the mean.
* 95% of the observations lie within two standard deviations of the mean.
* 99.9% of the observations lie within three standard deviations of the mean.

So far, we have been using “X” to denote the variable of interest (e.g., X=height, X=Weight). However, when using a standard normal distribution, we use “Z” to refer to a variable in the context of a standard normal distribution.

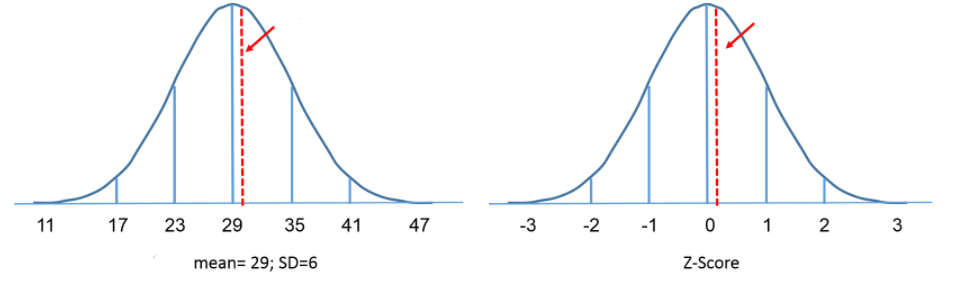
Example: We were looking at weight in a population of 15-year-old kids, who had normally distributed body weights with a mean value = 29 and a standard deviation = 6.

What is the probability that a randomly selected kid from this population will have a weight less than 30? While a value of 30 doesn’t fall on one of the increments of standard deviation. How can we calculate such probability using standard normal distributions?

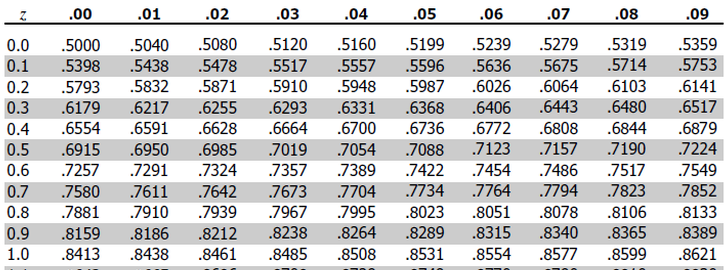
Solution: To compute P (X < 30) we convert the X=30 to its corresponding Z score (this is called standardizing):

Z = 30 – 29 /6 => 1/6

Z = 0.1667



Since the area under the standard curve = 1, we can begin to define the probabilities of occurrence of specific observation more precisely. The table in the frame below shows the probabilities for the standard normal distribution. Investigate the table and note that a “Z” score of 0.0 lists a probability of 0.50 or 50% and a “Z” score of 1, as one standard deviation above the mean, lists a probability of 0.8413 or 84%. This is because one standard deviation above and below the mean encompasses about 68% of the area, so one standard deviation above the mean represents half of that of 34%. Therefore, the 50% below the mean plus the 34% above the mean gives us 84%.



Z-table: <http://www.z-table.com/>

The Z value is the number of units of standard deviation which are away from the mean, and the area is the probability of observing a value less than that particular Z value. Also note that the table shows probabilities to two decimal places of Z. The units’ place and the first decimal place are shown in the left-hand column, and the second decimal place is displayed across the top row.

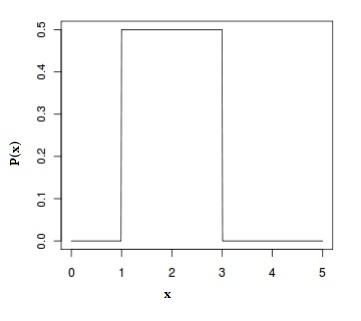
But let’s get back to the question about the probability that the weight is less than 30 (P (X < 30)).

Thus, P (X < 30) = P (Z < 0.1667). We can then look up the corresponding probability for this Z score from the table for standard normal distribution, which shows that P(X < 30) = P (Z < 0.1667) = 0.5636. Thus, the probability that a kid aged 15 has a body weight less than 30 is 56.36%.

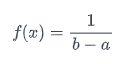
**Uniform Distribution / Rectangular Distribution:**

A uniform distribution, also called a rectangular distribution, is a probability distribution that has a constant probability. It has the following characteristics,

* This distribution is defined by **two parameters**, a and b:
* “a” is the minimum number.
* “b” is the maximum number.
* The distribution is written as U (a, b).
* The following graph shows the rectangular distribution with a=1 and b=3:

****

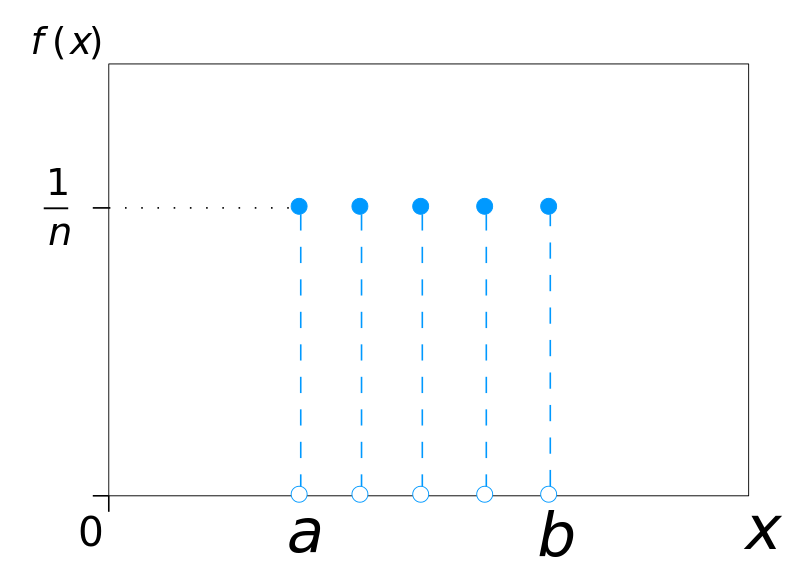
**Probability density function:**



Where:

* The length of the base of the rectangle is (b-a),
* The length of the height of the rectangle is 1/ (b-a).
* The area under f(x) and between the endpoints a and b is 1

This distribution has two types. The most common type used in [elementary statistics](https://www.statisticshowto.datasciencecentral.com/what-is-elementary-statistics/) is the **continuous uniform distribution** (shown as a rectangle). A second type called the **discrete uniform distribution** also resembles a rectangle but instead of a line, a series of dots represent a known, finite number of outcomes. The following graph shows 5 possible outcomes:



**Continuity Correction Factor:**

A continuity correction factor is applied when you use a continuous probability distribution to approximate a discrete probability distribution. For example: when you want to use the normal distribution to approximate a binomial distribution, following conditions need to be satisfied.

(n \* p) and (n \* q) >= 5

Where:

* n = how many items are in your sample.
* p = probability of an event happening.
* q = probability of an event not happening.

It is as simple as adding or subtracting 0.5 to the discrete x-value. Accordingly, use the following table to decide whether to add or subtract.

**Continuity Correction Factor Table:**

* if P (X = n) use P (n - 0.5 < X < n + 0.5)
* if P (X > n) use P(X > n + 0.5)
* If   P(X ≤ n) use    P(X < n + 0.5)
* If    P (X < n) use   P(X < n – 0.5)
* If    P(X ≥ n) use   P(X > n – 0.5)

Example: If n = 20 and p = .25, what is the probability that X >= 8?

**Step 1:** Start by working out n \* p and n \* q >= 5

np = 20 \* .25 => 5 (note: this is also the mean x̄)

nq = 20 \* .75 => 15

Both conditions are satisfied. Thus, we can use the Continuity Correction Factor.

**Step 2:** Find the variance of the binomial distribution:

n\*p\*q = 20 \* .25 \* .75 => 3.75

**Step 3:** Apply the Continuity Correction Factor on the X value. For this example, we have a greater than or equals sign (>=), so the table tells us:

P (X>n) use P (X > n – 0.5)

X >= 8 becomes X >= 7.5

**Step 4:** Find the z-score. You will need all three values from above:

* The mean from step 1.
* The variance from step 2.
* The Xi value from step 3.

Zi = 7.5 – 5 / √3.75 => 1.29

**Step 5:** Look up z value in the z-table (<http://www.z-table.com/>)

1.29 = 0.4015

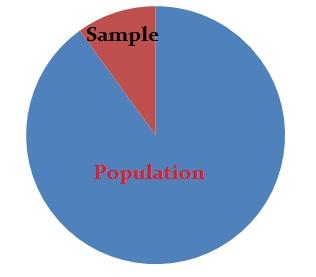
**Step 6:** Subtract probability from 0.5.

0.5 - 0.4015 = 0.0985

The probability that X >= 8 is 0.0985.



Example: To measure the diameter of each nail that is manufactured in a mill is impractical. Instead, you measure the diameters of a representative random sample of nails. Next, you use the information from the sample to generalize about the diameters of all of the nails.



## Sampling Distribution

The core goal of inferential statistics is to make intelligent conclusions about the population parameters by looking at the sample statistics.

Example: Estimate the mean height of the students in a class, from a small representative sample.

## Sampling Distribution of the Means

The sampling distribution of means is what you get if you consider all the possible samples of size “**n“,** taken from the same population and then form a distribution of their means. Each randomly selected sample is an independent observation.



This fact holds especially true for sample sizes which are greater than 30.

Visual demonstration: <http://onlinestatbook.com/stat_sim/sampling_dist/>

## **Expectation and Standard deviation of 𝑋** **bar**

Mean of all sample means of size n is the mean of the population.



Standard deviation tells us how far away from the population mean the sample mean is likely to be and is called the **Standard Error of the Mean**, and is represented as follows





## Two Main areas of inferential statistics

* Estimating Parameters.
* Test of Hypothesis.

One goal of statistical analyses is to obtain estimates of the population parameters and the amount of error associated with these estimates. These estimates are also known as sample statistics.

Example: Sample means are used to estimate the population means:

**There are two types of Parameter Estimates:**

* Point Estimate
* Confidence Intervals



Example: Xbar is a point estimate of the population mean μ. Similarly, the population proportion “p” is a point estimate of the population proportion “P”.

When we use samples to provide population estimates, we cannot be certain that they will be an accurate representation of the whole population.

For an example of parameter estimates, suppose you work for a bolt manufacturer that is studying a problem of the variations in diameters of bolts. It would be too costly to measure the diameter of every bolt in the company. Instead, you sample 100 randomly selected bolts and measure the diameter in millimeters. The mean of the sample is 9.2. This is the point estimate for the population mean (μ). You also create a 95% confidence interval for μ which is (8.8, 9.6). This means that you can be 95% confident that the true value of the average diameter of the bolt is between 8.8 and 9.6 mm.



Statisticians use the confidence interval to express the precision and uncertainty associated with a particular sampling method.

A confidence interval is defined as a range of values, derived from the sample statistics that is likely to contain the value of an unknown population parameter. Because of their random nature, it is unlikely that two samples from a particular population will yield identical confidence intervals. But if you repeat your sample several times, a certain percentage of the resulting confidence intervals would contain the unknown population parameter.

The interval estimate of a confidence interval is defined by:

**Confidence Interval = Sample Statistic + Margin of error**

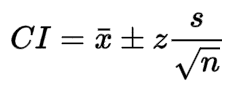
Where:

* Sample statistic for the mean is Xbar.
* Margin of Error is Z score \* SE (Standard Error).

Margin of error is defined as the maximum expected difference between the true population parameter and a sample estimate of that parameter.

## Confidence Intervals of Mean, Proportion, and Variance

## Mean:



Where:

* Xbar: Sample Statistics.
* Z: Z Score.
* S/square root(n) = Standard Error.

Example:

A survey involving US companies that trade business with firms in India was taken. One of the survey questions was: Approximately how many years has your company been trading with firms in India?

A random sample of 44 responses to this question yielded a mean of 10.455 years. Suppose that the population standard deviation for this question is 7.7 years. Using this information, construct a 90% confidence interval for the mean number of years that a certain company has been trading with firms in India.

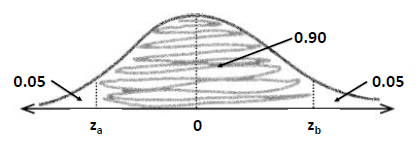
Solution:

* n = 44
* Xbar = 10.455
* σ = 7.7

Confidence interval for the population mean is with 90% probability.

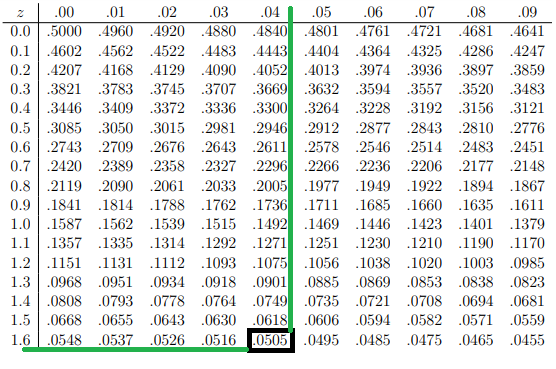
Sample mean ± Margin of Error

Find Za and Zb where P (Za < Z < Zb) = 0.90



P (Z < Za) =0.05 and P (Z > Zb) = 0.05

From Z table: (<http://www2.stat.duke.edu/~rcs46/lectures_2015/14-bayes1/z-table.pdf>)



Therefore, Z value for 5% probability is 1.64.

Za = -1.64 and Zb = 1.64

Margin of error at 90% CI = 1.64 \* (7.7/sqrt (44)) = 1.91

(Xbar – 1.91) < 𝜇 < (Xbar + 1.91)

Since the sample mean is 10.455 years, we get the confidence interval for 90% as

8.545 < 𝜇 < 12.365

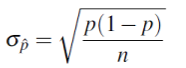
The analyst is 90% confident that if a census of all the US companies trading with firms in India were taken at the time of the survey, the actual population mean number of trading years of such firms would be between 8.545 and 12.365 years.

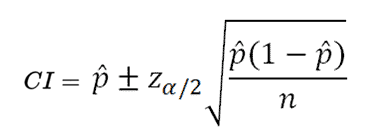
## Proportion:

The population proportion is denoted p and the sample proportion is denoted pˆ, Thus if 43% of people entering a store make a purchase, p = 0.43; if in a sample of 200 people entering the store, 78 make a purchase, pˆ = 78/200 => pˆ = 0.39.

Suppose the random samples of size n are drawn from a population in which the proportion with a characteristic of interest is p. Then mean μP and standard deviation σP^ of the sample proportion   
P^ satisfy,



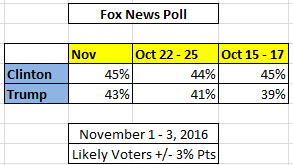




Where:

* P = sample proportion
* Square root (p(1-p)/n) = Standard Error of sample proportion
* Z = Z score

**Example**: In a poll by FOX News conducted between November 1 and 3, 2016, a survey of 1107 randomly sampled likely voters predicted that 45% would vote for Hillary Clinton:



Then what is the margin of error at 95% confidence level (Z = 1.96)?

Solution:

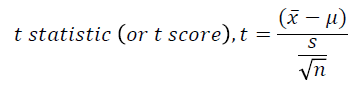
* n = 1107
* p = 0.45
* q = 0.55
* z = 1.96

Margin of error = 1.96 \* (sqrt ((0.45\*0.55)/1107))

**Margin of error = 2.93%**

## t-Distribution:

If the **sample size is small (<30)**, the variance of the population is not adequately captured by the variance of the sample. In such cases, instead of z-distribution, t-distribution is used. It is also the appropriate distribution type to use when the **population variance is not known**.



Degrees of freedom (v) is defined as the number of independent observations for a source of variation minus the number of independent parameters estimated in computing the variation.

When estimating the mean of the population from a single sample, the number of independent observations is equal to n-1.

Properties of t-Distribution are as follows,

* Mean of the distribution = 0.
* Variance = v/ (v-2), where v > 2.
* Variance is always greater than 1; although it is close to 1, when there are multiple degrees of freedom (sample size is large).
* With infinite degrees of freedom, t-distribution is the same as the standard normal distribution.

Confidence Interval to Estimate μ:



* Population standard deviation is unknown, and the population is normally distributed.
* Sample mean, standard deviation, and size can be calculated from the data; t value can be inferred from the table.
* α is the area in the tail of the distribution. For 90% confidence level, α= 0.10. In a confidence Interval, this area is symmetrically distributed between 2 tails (α/2 in each tail).

Example:

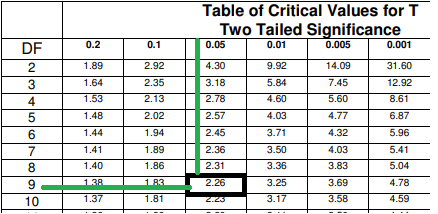
The labeled potency of the dosage for a tablet is 100mg. As per the quality control specifications, 10 tablets are randomly assayed for the potency.

A researcher wants to estimate the interval for the true mean of the batch of tablets with 95% confidence. Assume that the potency is normally distributed.

Data are as follows (in mg):

| 98.6 | 102.1 | 100.7 | 102 | 97 |
| --- | --- | --- | --- | --- |
| 103.4 | 98.9 | 101.6 | 102.9 | 105.2 |

* Mean, Xbar = 101.24 mg
* Standard deviation, S = 2.48
* n = 10
* degree of freedom = 10 -1 = 9
* At 95% level, α = 0.05, and ∴, 𝛼/2=0.025
* t two tailed table ([http://snobear.colorado.edu/Markw//IntroHydro/12/statistics/testchart.pdf](http://snobear.colorado.edu/Markw/IntroHydro/12/statistics/testchart.pdf))



**t value from the table is 2.26**



101.24 – (2.62 \* (2.48/sqrt (10)) ≤ μ ≤ 101.24 + (2.62 \* (2.48/sqrt (10))

99.47 ≤ μ ≤ 103.01

The mean value for the batch is 101.24mg with an error of +/-1.77 mg. The researcher is 95% confident that the average potency of the batch of tablets is between 99.47 mg and 103.01 mg.

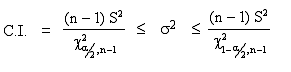
## Chi Square Distribution for Variance:

A standard normal distribution with **k** degrees of freedom is defined as the distribution of a sum of the squares of **k** independent of standard normal random variables. It is widely used in inferential statistics, notably in deriving confidence interval and hypothesis testing.

Chi Square distribution is positively skewed, with the degree of skew decreasing with increasing degree of freedom. Accordingly, as the degrees of freedom increases, the Chi Square distribution approaches the normal distribution.

You can use the Chi Square distribution to construct the confidence interval for the variance and standard deviation.

Formula:

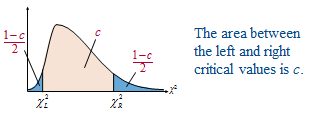


Where:

* S = Sample standard deviation.
* n = sample size.
* χ² = Chi Square value from Chi Square table.
* All χ² values are greater than or equal to zero.

Critical Values for χ²:

* There are two critical values for each level of confidence.
* The values *χ*2R represent the right-tail critical value.
* The values *χ*2L represent the left-tail critical value.

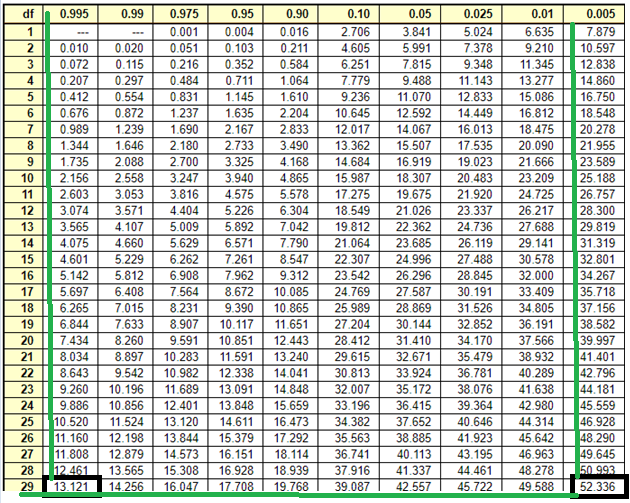


Example:

You randomly select and weigh 30 samples of an anti-allergy medicine. The standard deviation for the samples is 1.20 milligrams. Assuming that the weights are normally distributed, construct a 99% confidence interval for the population variance.

Solution:

* n = 30
* df = n -1 = 30 – 1 = 29
* Area to the right of *χ*2R = (1 – c)/2 = (1 – 0.99)/2 = 0.005
* Area to the left of *χ*2L = (1 + c)/2 = (1 + 0.99)/2 = 0.995
* The critical values are *χ*2R = 52.336 and *χ*2L = 13.121



*χ*2 table (<https://people.richland.edu/james/lecture/m170/tbl-chi.html>)

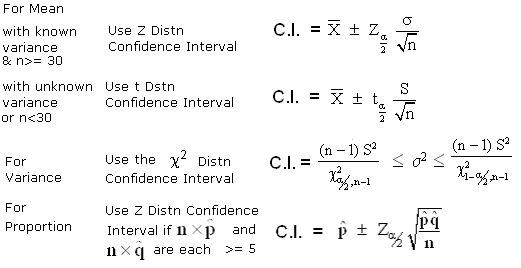
Confidence Interval for **σ2**:

* Left endpoint: (n – 1)S2/ *χ*2R = (30-1)(1.20)2 / (52.336) = 0.80
* Right endpoint: (n – 1)S2/ *χ*2L = (30-1)(1.20)2 / (13.121) = 3.18

0.80 < **σ2** < 3.18

You can conclude with 99% confidence that the population variance is between 0.80 and 3.18 milligrams.

## Confidence Intervals Formulas



Example:

A school principal claims that the students from her school have an average score of 7/10 in the English Proficiency test. You doubt the claim, take a random sample of 40 students for analysis and find a mean score of 5.5/10, with a sample standard deviation of 1. Can you reject the principal’s claim?

**Step 1:** Decide on the hypothesis.

Average score on the test is 7/10.

This is called **Null Hypothesis** and is represented by H0.

In this case,

**H0: 𝜇 = 0.7**

If the Null Hypothesis is rejected based on evidence, an alternate Hypothesis, H1, needs to be accepted. We always start with the assumption that Null Hypothesis is true.

In this case,

**H1: 𝜇 < 0.7**

**Step 2:** Choose your statistic,

Sample size = 40

Normal distribution is a good approximation,

Standard Error = s / sqrt (n) = 1.0 / sqrt (40)

Standard Error = 0.158

X ~ N (0.7, 0.1582) = N (0.7, 0.025)



Z = (0.55 – 0.7) / 0.158

**Z = 0.94**

**Step 3:** Specify the Significance Level.

First, we must decide on the Significance Level, α, which is defined as the measure of how unlikely you want the results of the sample to be before you reject the null hypothesis, H0.

**Step 4:** Determine the critical region.

If X represents the mean score of the sample, the critical region is defined as P (X < c) < α where α=5%



Recall that in a 95% Confidence Interval (CI), there is a 5% chance that the sample will not contain the population mean. Hence if the sample falls in the critical region, the null hypothesis that 0.7 is the mean score is rejected.

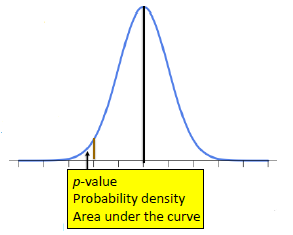
This is the reason that 5% or 0.05 is called the Significance Level. In a 99% CI, 0.01 is the Significance Level.

**Step 5:** Find the *P*-value

P-value is the probability of getting a value up to and including the one in the sample in the direction of the critical region.

It is a way of taking the sample and working out whether the results fall within the critical region of the hypothesis test.

It indicates how likely it is that a result was obtained by chance alone. Essentially, this is the value used to determine whether to reject the null hypothesis.



If the p-value is small, it indicates the result is unlikely to have occurred by chance alone. These results are known as being statistically-significant results.

* P-value ranges from 0 to 1.
* Small p-value = greater than chance alone (test is significant).
* Large p-value = result is within chance (test is not significant).

If α = 0.05 or 5%, the following rules apply:

* If p < α, the test is statistically significant.
* If p > α, the test is statistically not significant.

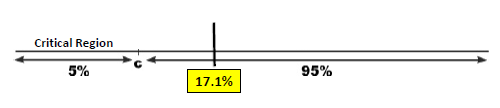
In our sample, we found a mean score of 5.5/10. This implies that our p-value is P(X ≤ 0.55), where X is the distribution of the mean scores in the sample.

If P (X ≤ 0.55) < 0.05 (Significance Level), it indicates that 0.55 is inside the critical region, and hence H0 can be rejected.

Given that Z = -0.94, P (X ≤ 0.55) = 0.171 

Thus, there is a 17% probability of finding a mean score of 5.5/10 or less.

**Step 6:** Is the sample result in the critical region?



The sample result is not in the critical region.

**Step 7:** Make your decision.

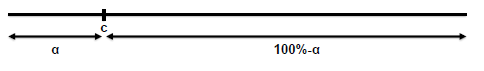
There isn’t sufficient evidence to reject the null hypothesis and therefore, the claim of the principal is accepted.

## Critical Region up Close

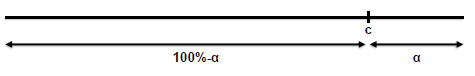
## One-tailed tests:

The position of the tail is dependent on H1 (Alternate Hypothesis).

If H1 includes a **“<” sign (lesser than)**, then the lower tail is utilized.

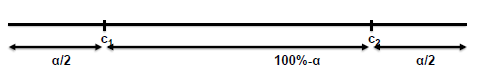


If H1 includes a **“>” sign (greater than)**, then the upper tail is utilized.



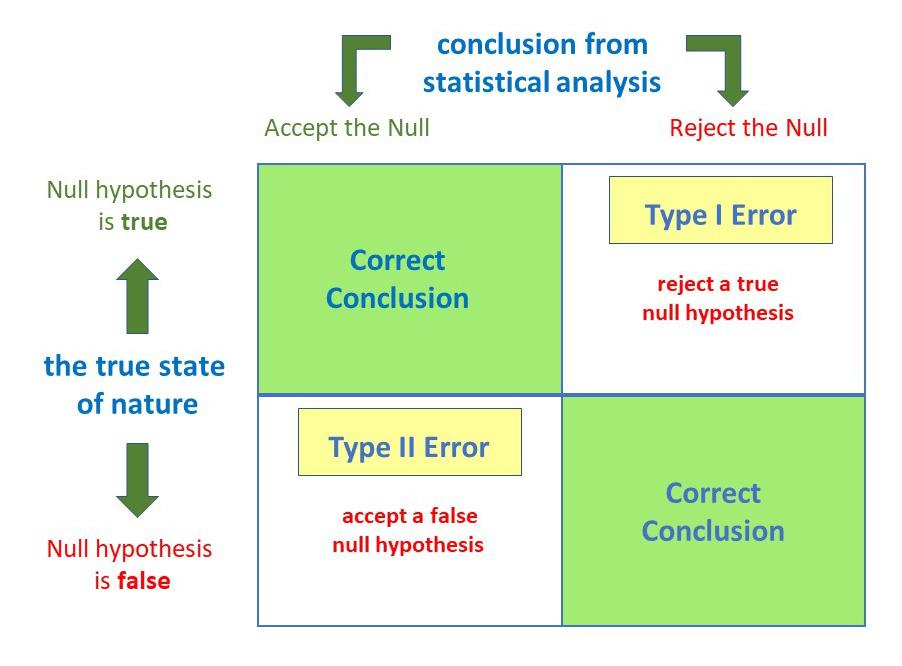
## Two-tailed tests:

Critical region is split over both ends. Both ends contain α/2, making a total of α. If H1 includes a **“≠” sign (Not equal to)**, then the two-tailed test is used, since we then look for a change in parameter, rather than an increase or a decrease.



## Types of Errors

* Type I: We *reject* the NULL Hypothesis incorrectly.
* Type II: We “*accept*” the NULL Hypothesis incorrectly.



Source: <https://www.simplypsychology.org/type-1-and-2-errors.jpg>

## Type I error:

A type I error is also known as the false positive error and occurs when a researcher incorrectly rejects a true null hypothesis. This means that you report that your findings are significant when in fact they have occurred by chance and are statistically not significant.

The probability of making a type I error is represented by your alpha level (α), which is the p-value below which you reject the null hypothesis. A p-value of 0.05 indicates that you are willing to accept a 5% chance that you are wrong in case the null hypothesis is rejected.

You can reduce the risk of committing a type I error by using a lower value for p. For example: a p-value of 0.01 would mean there is a 1% chance of committing a Type I error.

However, using a lower value for the alpha means that you will be less likely to detect a true difference, if one really exists (thus risking a type II error).

## Type II error:

A type II error is also known as false negative and occurs when a researcher fails to reject a null hypothesis which was false. Here a researcher concludes that there is no significant effect, when there really is.

The probability of making a type II error is called Beta (β), and this is related to the power of the statistical test (power = 1- β). You can decrease your risk of committing a type II error by ensuring that your test has enough power.

You can do this by keeping your sample size large enough to detect a practical difference, which truly exists.

## Why are Type I and Type II Errors Important?

The consequences of making a type I error mean that unnecessary changes or interventions are made, which results in wastage of time and resources, etc.

Type II errors typically lead to the preservation of the status quo (i.e. interventions remain the same) when in reality a change is required.

Example:

Scientists claim that a new miracle drug cures the common cold with a success rate of 90%. You conduct a random sample test on 100 patients and find that 80 of them are cured.

At 5% significance level, do you reject or accept the claim by the scientists?

What are the Null and Alternative Hypotheses?

H0: p = 0.9

H1: p < 0.9

What is the test statistic?

X ~ B (100, 0.9)

Since np>5 and nq>5, Normal distribution can be used instead.

X ~ N (np, npq)

X ~ N (90, 9)

What is the probability of 80 or fewer patients getting cured?

Z = (80.5 – 90) / sqrt (9)

**Z = -3.17**

P-value = P (Z < -3.17)

**P-value = 0.0008**

## Probabilities of Errors in our Example:

Probability of Type I error

P (Type I error) = 0.05

To calculate P (Type II error)

H0: p = 0.9

H1: p = 0.8

P (Z < c) = 0.05 for 5% Significance Value.

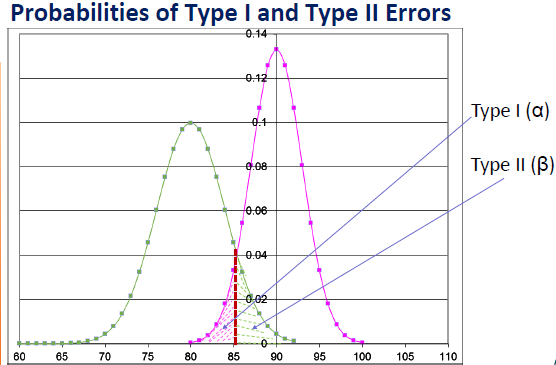
From probability tables, c = -1.64

To de-standardize and find values outside the critical region,

X – 90 / sqrt (9) ≥ -1.64

X ≥ 85.08

i.e., we will accept the null hypothesis if 85.08 or more people had been cured.



Finally, we need to calculate P(X ≥85.08), assuming H1 is true. X ~ N (np, npq) where n=100 and p=0.8. This gives X ~ N (80, 16). To calculate P(X ≥85.08) where X ~ N (80, 16), we find:

Z = (85.08 – 80) / sqrt (16)

Z = 1.27

P (Z ≥1.27) = 1 – P (Z < 1.27) = 1-0.8980 = 0.102

P (Type II error) = 0.102

The probability of accepting the null hypothesis that 90% of patients are cured when only 80% are cured in reality is 10.2%.

## Power of Hypothesis Test:

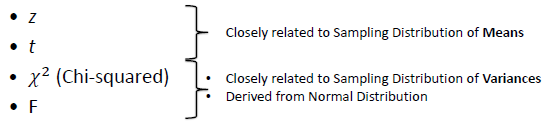
We reject the null hypothesis correctly when it is false. It is the opposite of Type II error, and therefore,

Power = 1 –β= 1-0.102 = 0.898, i.e.

The probability that we will take the correct decision in rejecting the Null Hypothesis is 89.8%

## Common Test Statistics for Inferential Techniques

Inferential techniques (Confidence Intervals and Hypothesis Testing) most commonly use 4 test statistical methods:



## Types of t – tests

There are three main types of t-tests:



## Two-Sample t-test for Paired Data:

When the effects of two alternative treatments are compared, it is sometimes possible to make comparisons in pairs, where, e.g. the pair can be the same person at two different time-points or matched pairs where they are alike in all respects.

To study if their mean values are the same – we can create a new data set from the difference of the individual data points.

**Xnew = X1 – X2**

Subsequently, we can look at how far away from zero is the mean E (Xnew)



Example:

A Yoga trainer suggests that meditation increases concentration. To test this hypothesis, you get 12 volunteers and note the time it takes to complete a puzzle by them. The next day, you put them through a 30 minute meditation routine and ask them to complete another puzzle of similar difficulty. The time taken for completion is measured again.

You want to test at 5% Significance Level (or 95% Confidence Level) if the time taken is shorter after meditation.

| Time to Solve the puzzle (Minutes) | | | |
| --- | --- | --- | --- |
| Volume | After | Before | A-B |
| 1 | 63 | 55 | 8 |
| 2 | 54 | 62 | -8 |
| 3 | 79 | 108 | -29 |
| 4 | 68 | 77 | -9 |
| 5 | 87 | 83 | 4 |
| 6 | 84 | 78 | 6 |
| 7 | 92 | 79 | 13 |
| 8 | 57 | 94 | -37 |
| 9 | 66 | 69 | -3 |
| 10 | 53 | 66 | -13 |
| 11 | 76 | 72 | 4 |
| 12 | 63 | 77 | -14 |
| Total | **842** | **920** | **-78** |
| Mean | **70.17** | **76.67** | **-6.50** |

Mean of the difference,

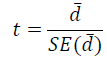


Standard Deviation of the differences,



Standard Error of the mean,





t = - 6.5 / 4.37

**t = -1.487**

Number of degrees of freedom = 12 -1 = 11

Framing the hypothesis

H0: µ1 = µ2 (before = after)

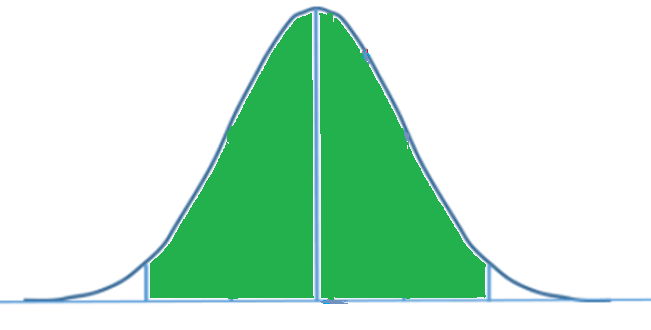
H1: µ1 ≠µ2 (before ≠ after)

Since H1 includes a **“≠” sign (Not equal to)**, we use the two-tailed test.

**t = -1.487**

**t11, 0.05 = 1.7958**

(From t table)



Comparing the absolute t-value, we **cannot reject** the Null Hypothesis which means that the completion time taken by two tests is the same.

The 95% CI for the mean difference is given by,



-6.5 – (1.7958 \* 4.37) ≤ D ≤ -6.5 + (1.7958 \* 4.37)

**95% CI: (14.34, 1.34)**

Since the value zero is included in the CI, we cannot reject the null hypothesis.

**Business Decision**

Although zero is included in CI, the range is very wide, which demands conducting a larger study to confidently assert our conclusion.

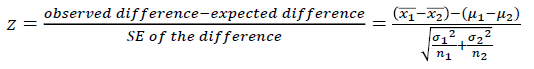
## Two-Sample t-Test for Unpaired Data:

The Central Limit Theorem states that the difference between the two sample means, (X1bar – X2bar) is normally distributed for large sample sizes (both n1 and n2 ≥ 30) irrespective of the population distribution.

Also,





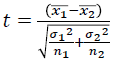


This is the test statistic for a 2-sample z-test.

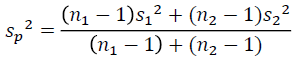
H0: µ1 = µ2

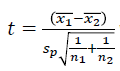
H1: µ1 ≠µ2

t-test statistic,



Assuming that the two samples come from the populations with the **Same Standard Deviation** (Rule of thumb: The ratio between the higher S and the lower S is less than 2), pooled variance can be used to calculate SE.





With (n1 and n2 – 2) degrees of freedom

**Example**:

Antibiotic rifampicin increases the amount of drug-metabolizing enzymes present in the liver. This causes an increase in the rate of elimination of a number of drugs.

An experiment was conducted to study whether rifampicin affects the metabolic removal of the anti-asthma drug, theophylline. A high elimination rate would mean an inadequate treatment of the patient’s asthma.

Two groups of 15 subjects were first with oral rifampicin (600 mg once daily for 10 days) and a placebo drug, respectively. They were then given intravenous injection of theophylline (3 mg/kg of body weight).

Drug levels were measured from the blood samples and efficiency of removal of theophylline was reported as clearance (in ml/min/kg).

| **Clearance of theophylline** | |
| --- | --- |
| **Control Group** | **Treated Group** |
| 0.81 | 1.15 |
| 1.06 | 1.28 |
| 0.43 | 1.00 |
| 0.54 | 0.95 |
| 0.68 | 1.06 |
| 0.56 | 1.15 |
| 0.45 | 0.72 |
| 0.88 | 0.79 |
| 0.73 | 0.67 |
| 0.43 | 1.21 |
| 0.46 | 0.92 |
| 0.43 | 0.67 |
| 0.37 | 0.76 |
| 0.73 | 0.82 |
| 0.93 | 0.82 |

**n1 = 15, X1bar = 0.931, S1 = 0.202, S12 = 0.0408**

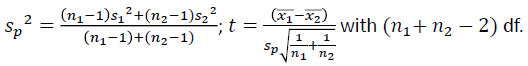
**n2 = 15, X2bar = 0.633, S2 = 0.216, S22 = 0.0467**

Framing the Hypothesis:

H0: µ1 - µ2 = 0 (Rifampicin does not cause a change in theophylline clearance)

H1: µ1 - µ2 ≠ 0(Rifampicin does cause a change in theophylline clearance)

It is a Two-tailed test.

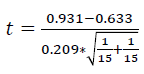


Sp2 = (((15-1) \* (0.0408)) + ((15-1) \* (0.0467))

(15 - 1) \* (15 + 1)

Sp2 = 0.04375

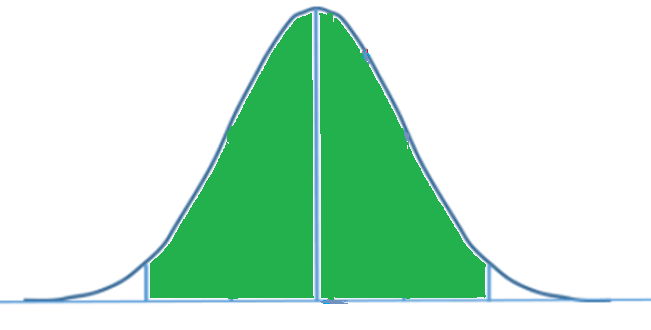
Sp = 0.209



t = 3.91

You can find the p-value for this t-score by knowing that the t-score is much larger than the critical value for 28 degrees of freedom (~2) at 5% significance level,

t28, 0.05 = 2.05



You see that the p value is in the critical region in the right tail.

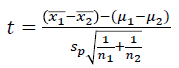
Will you reject the null hypothesis or fail to do so?

**Reject**. That means rifampicin does affect the clearance of theophylline.

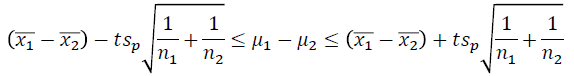
Does rifampicin increase or decrease the theophylline clearance and by what degree?

As the treated patients exhibited a higher clearance (0.931 ml/min/kg) compared to the control group (0.633 ml/min/kg), rifampicin increases clearance by about 0.298 ml/min/kg).

## Confidence Intervals



Rewriting



((0.298) – (2.048 \* (0.0763)) ≤ µ1 - µ2 ≤ ((0.298) + (2.48 \* (0.0763))

95% CI: (0.142, 0.454)

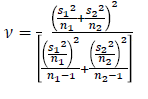
Note that the zero difference is unlikely at 95% Confidence Level, as the difference ranges between 0.142 and 0.454 ml/min/kg; with a point estimate for the difference in the mean clearance being 0.298 ml/min/kg.

**Welch’s t-test using Welch-Satterthwaite equation to calculate the degrees of freedom**



When two standard deviations are not equal, we use the above formula.

In this case, the degree of freedom is calculated as:



Round off to the nearest integer.

## AIM

The aim of the following lab exercises is to perform partial implementation of the Normal distribution, Standard Normal distribution, Confidence Intervals, and Test of Hypothesis by writing python code so that we get hands-on practice of Inferential Statistics.

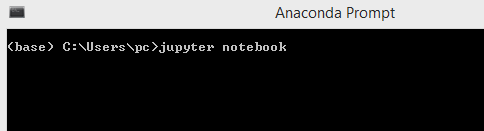
The labs for this chapter include the following exercises.

* Calculating the probability using standard normal distribution.
* Calculating the confidence intervals.
* Performing test of hypothesis.

We will be working with Python3 and jupyter notebook IDE.

**Launching Jupyter notebook**

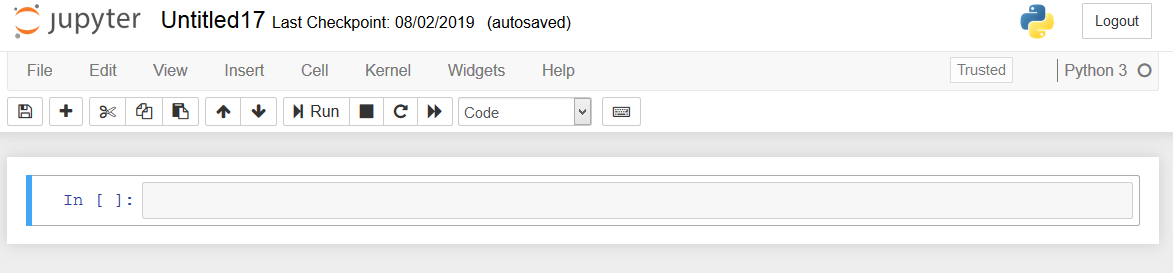
To start THE notebook server, enter the following command in your Anaconda Prompt.



After starting the notebook server, create a new notebook by clicking on new and python3.

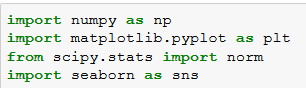


Now we have our new notebook, we can start with implementation.



## Importing the necessary packages

We always need to import the necessary packages required for the project or task.



## Calculating probabilities using normal and standard normal distributions

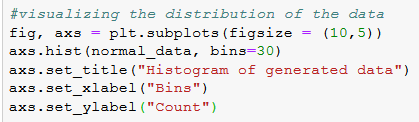
**Task 1:** Generate the data that follows the normal distribution and visualize the distribution.

Step 1: We are generating 10000 data points, which follow the normal distribution with mean = 0 and standard deviation = 1.

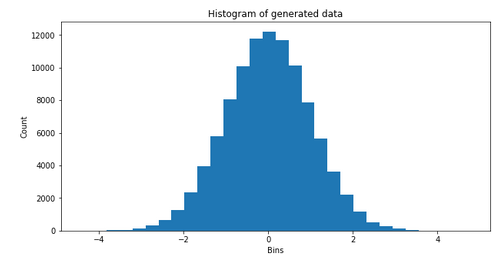


The function “np.random.randn” samples 10000 data points from the normal distribution which has mean = 0 and standard deviation = 1.

Step 2: Visualizing the distribution of the data generated.



* In line 1, we set the size of the figure to be (10, 5).
* In line 2, we plotted a histogram with the data we generated.
* Line 3 – 5, we formatted the figure with titles, x-label and y-label.



We can clearly see that the histogram is bell-shaped with a mean value of = 0.

**Task 2:** Generate a normal data set with mean = 65 and standard deviation = 3, and visualize the distribution.

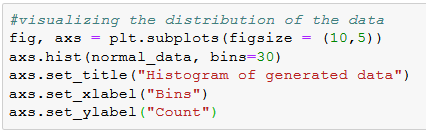
Step 1: In the previous task, we generated a normal data set with default mean and standard deviation values. Here in this task, we will generate a normal data set with mean = 65 and standard deviation = 3.

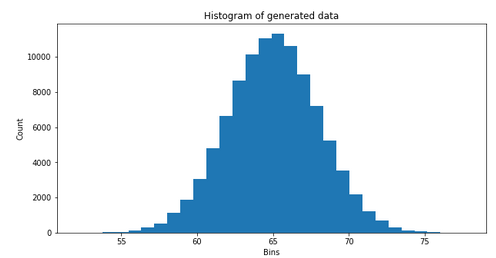


The function “np.random.normal” helps you to generate a normal data set with custom mean and standard deviation values. The function takes three important parameters into consideration:

* loc: The location where the mean of data needs to be.
* scale: It indicates the standard deviation of the data.
* size: The size parameter tells how many data points to be sampled from the normal distribution.

Step 2: Visualizing the distribution





We can clearly see that the data is normally distributed with mean = 65.

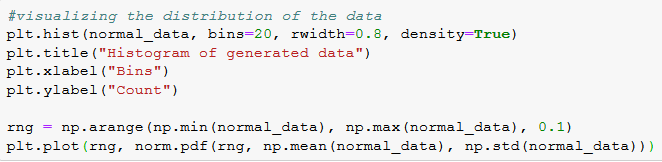
**Task 3:** Students’ scores in 1000 exams are normally distributed with mean values of 65 and a standard deviation of 9. Find the percent of the scores, which are:

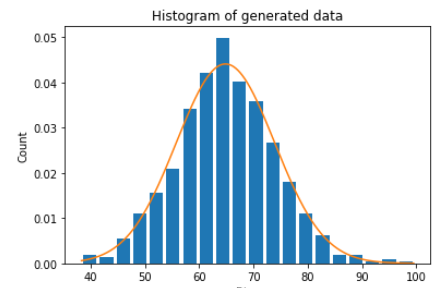
1. less than 55.
2. at least 85.
3. between 70 and 80.

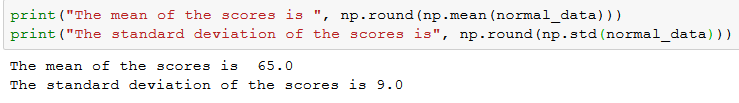
Step 1: Generate a score of 1000 exams, which follows the normal distribution with mean = 65 and standard deviation of 9.



Step 2: Visualize the distribution,







The scores of 1000 exams are normally distributed with mean = 65 and standard deviation = 9.

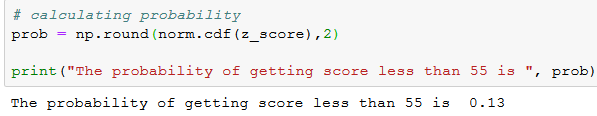
**Task 3:** Calculate the following probabilities:

1. What is the probability of getting an exam score less than 55, P (X < 55)?

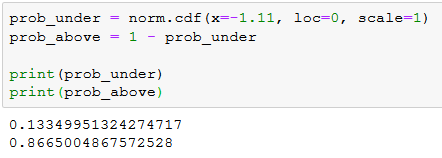
Step 1: Calculate the z-score for x = 55.

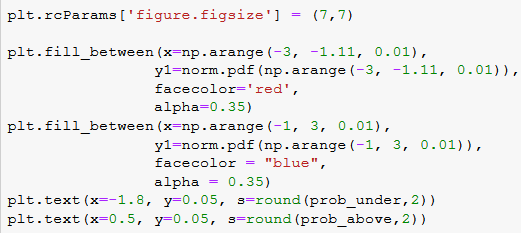


Step 2: Calculate the area under the curve,

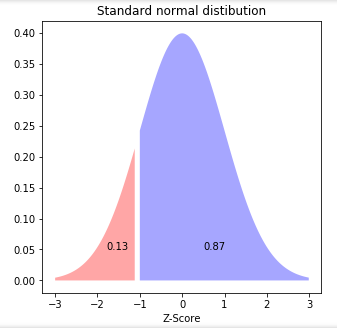


Step 3: Visualize and interpret the results,





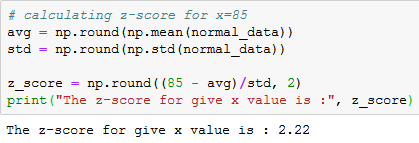
 Python calculates the left / lower-tail probabilities by default.



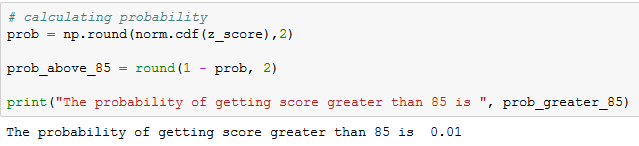
For this curve z = -1.11 and the area under left is 0.13. Therefore, the probability of getting an exam score less than 55 is 0.13; P (X < 55) is 13%.

1. What is the probability that the exam score will be at least 85; P(X ≥ 85)?

Step 1: Calculate the z-score for x = 85.

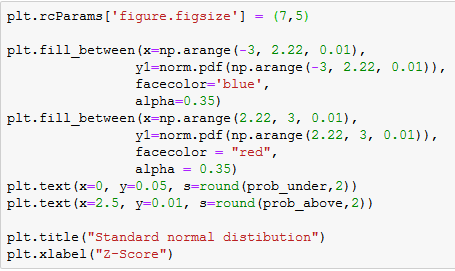


Step 2: Calculate the area under the curve.

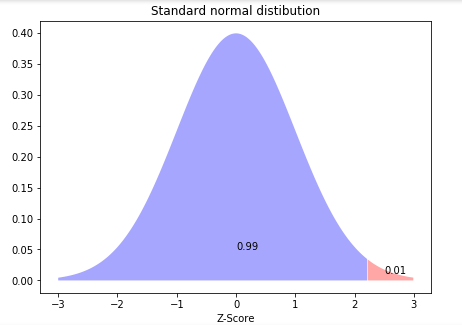


Step 3: Visualize and interpret the results.





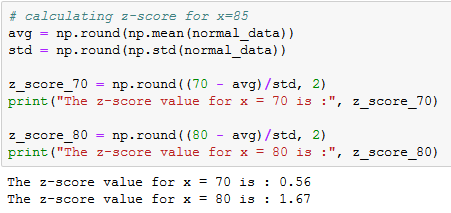
 Python calculates the left / lower-tail probabilities by default.



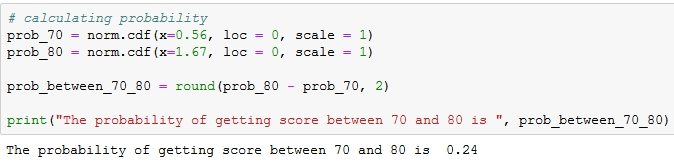
For this curve z = 2.22 and the area under right is 0.01. Therefore, the probability of getting an exam score greater than 85 is 0.01; P (X < 85) is 1%.

1. What is the probability that the exam score will be between 70 and 80; P (70 < X < 80)?

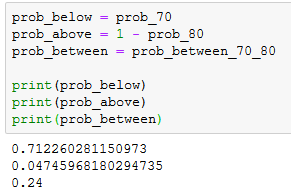
Step 1: Calculate the z-score for x = 70 and x = 80.

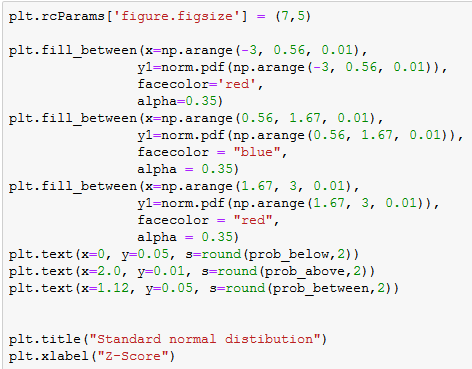


Step 2: Calculate the area under the curve,

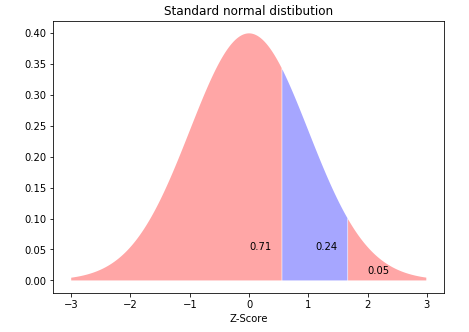


Step 3: Visualize and interpret the results.





 Python calculates the left / lower-tail probabilities by default.



For this curve, z values are 0.56 and 1.67. Therefore, the probability of getting an exam score between 70 and 80 is 0.24; P (70 < X < 80) is 24%.

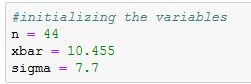
## Calculating the confidence intervals

**Computing the confidence interval for a population mean**

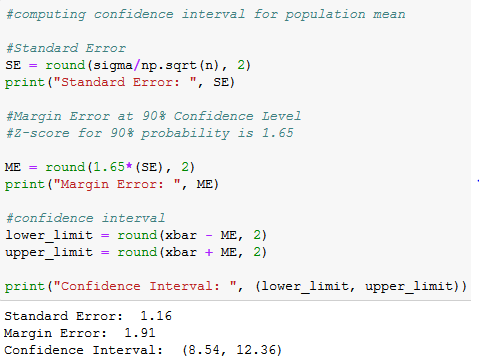
**Task 4:** A survey was conducted of the US companies that do business with many firms in India. One of the survey questions was: Approximately how many years have your company been trading with the firms in India?

A random sample of 44 responses to this question yielded a mean value of 10.455 years. Suppose the standard deviation for this population is 7.7 years. Use this information to construct a 90% confidence interval for the mean number of years that a company has been trading in India, out of the population of US companies trading with firms in India.

Step 1: Initialize the variables required for calculating the confidence interval.



Step 2: compute the standard error, margin error, and confidence intervals.



The analyst is 90% confident that if a census of all the US companies trading with the firms in India were taken at the time of the survey, the actual population mean number of trading years would be between 8.45 and 12.36 years.

**Calculating the confidence interval for a population mean using t-distribution.**

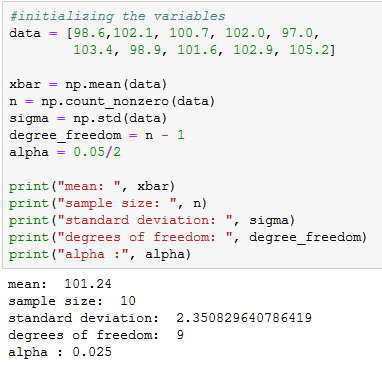
**Task 5:** The labeled potency of a medicinal tablet is 100mg. As per the quality control specifications, 10 tablets are randomly assayed for potency testing.

A researcher wants to estimate the interval for the true mean of the batch of tablets with 95% confidence. Assume that the potency is normally distributed.

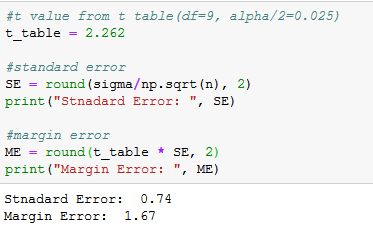
Data is as follows (in mg):

| 98.6 | 102.1 | 100.7 | 102 | 97 |
| --- | --- | --- | --- | --- |
| 103.4 | 98.9 | 101.6 | 102.9 | 105.2 |

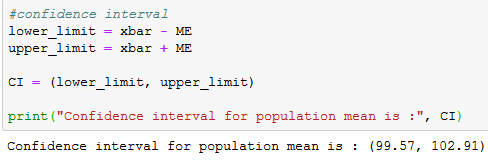
Step 1: Initialize the variables required for calculating the confidence interval.



Step 2: Calculate the standard error and margin error.



Step 3: Calculate the confidence interval and interpret it.

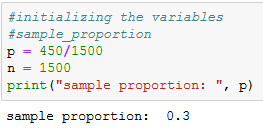


The mean for the batch is 101.24mg with an error of +/-1.67mg. The researcher is 95% confident that the average potency of the batch of tablets is between 99.57mg and 102.91mg.

**Calculate the confidence interval for a population proportion.**

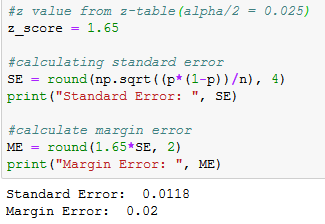
**Task 6:** Let’s suppose that on a certain website, out of 1500 visitors on a given day, 450 clicked on an ad purchased by a sponsor. Let’s construct a confidence interval for the population proportion of visitors who clicked on the ad.

Step 1: initialize the variables.

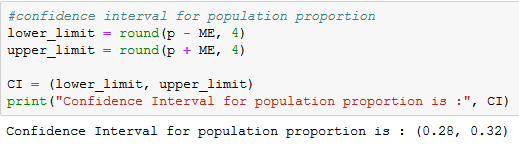


We have initialized the sample proportion and number of sample variables.

Step 2: Calculate the standard error and margin error.



Step 3: Compute the confidence interval for the population proportion.



We are 95% confident that the population proportion falls between 0.28 and 0.32.

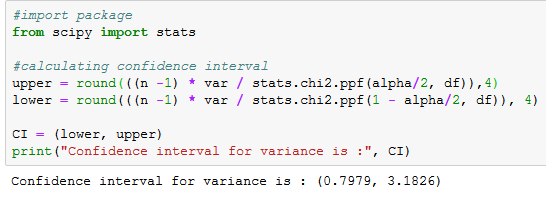
**Calculate the confidence interval for variance.**

**Task 7:** You randomly select and weigh 30 samples of an anti-allergy medicine. The sample standard deviation is 1.20 milligrams. Assuming that the weights are normally distributed, construct the 99% confidence interval for the population variance.

Step 1: initialize the variables.



Step 2: Calculating the confidence intervals.



We are 99% confidence that our population variance falls between 0.7979 and 3.1826.

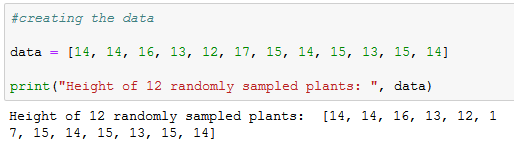
## Performing Test of Hypothesis

**One sample t-Test**

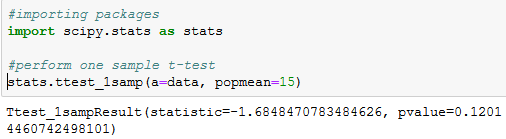
**Task 8:** Suppose a botanist wants to know if the mean height of a certain species of plant is equal to 15 inches. She collects a random sample of 12 plants and records their individual heights in inches.

Use the following steps to conduct a one sample t-test to determine if the mean height for this species of plant is equal to 15 inches.

Step 1: Create the data.



Step 2: Conduct a one sample t-test.



We used the “ttest\_1samp()” function from the “scipy.stats” package to conduct a one sample t-test, which uses the following syntax:

**ttest\_1samp(a, popmean)**

Where:

* a: an array of sample observations.
* popmean: the expected population mean.

The t-test statistic is -1.6848 and the corresponding two-sided p-value is 0.1201.

The two hypotheses for the one sample t-test are as follows:

H0: µ = 15 (The mean height for this species of plant is 15 inches).

H1: µ ≠ 15 (The mean height is not 15 inches).

Step 3: Interpret the results.

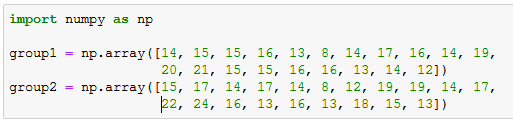
Because the p-value of our test (0.1201) is greater than alpha = 0.05, we fail to reject the null hypothesis. We do not have sufficient evidence to say that the mean height for this specie of plant is different than 15 inches.

**Two sample t-test**

**Task 9:** Researchers want to know whether two different species of plants have the same mean height. To test this, they collect a simple random sample of 20 plants from each species.

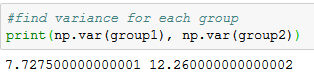
Use the following step to conduct a two-sample t-test to evaluate if any two species of plants have the same height.

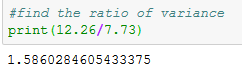
Step 1: Create the data



Step 2: Conduct a two-sample t-test.

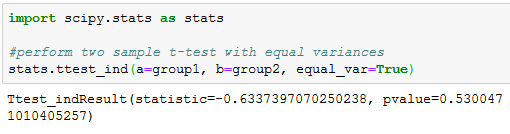
Before we perform the test, we need to decide if the two populations have equal variances or not. As a rule of thumb, we can assume that the populations have equal variance if the ratio of the larger sample variance to the smaller sample variance is less than 4:1.





The ratio of the larger sample variance to the smaller sample variance is 1.586, which is less than 4. Thus, we can assume that the population variances are equal.

Next, we proceed to perform the two-sample t-test with equal variances:



We have used the “ttest\_ind()” function from scipy.stats library to conduct a two-sample t-test, which uses the following syntax:

**ttest\_ind(a, b, equal\_var=True)**

Where:

* a: an array of the sample observations for group 1.
* b: an array of the sample observations for group 2.
* equal\_var: if True, perform a standard independent two sample t-test that assumes equal population variances. If False, perform “Welch’s t-test”, which does not assume the equal population variances. This is True by default.

The t-test statistic is -0.6337 and the corresponding two-sided p-value is 0.53005.

Step 3: Interpret the results.

The two hypotheses for this two sample t-test are as follows:

H0: µ1 = µ2 (The mean values for the two populations are equal).

H1: µ1 ≠ µ2 (The mean values for the two populations are not equal).

Because the p-value of our test (0.53005) is greater than alpha = 0.05, we fail to reject the null hypothesis. Thus, we do not have sufficient evidence to say that the mean height of plants between the two populations is different.

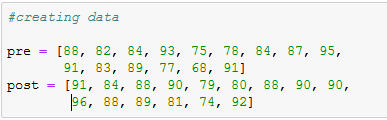
**Paired sample t-test**

**Task 10:** Suppose we want to know whether a certain study program significantly impacts the student performance in an exam. To test this, 15 students in a class take a pre-test. Then, the students participate in the study program for two weeks. Following this, the students retake a test of similar difficulty.

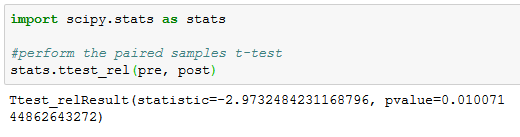
To compare the difference between the mean scores on the first and second test, we use a paired samples t-test. This is because the first test score can be paired with the second test score for each student.

Perform the following steps to conduct a paired sample t-test.

Step 1: Create the data.



Step 2: Conduct a paired samples t-test.



We used “ttest\_rel()” function from the ‘scipy.stats’ library to conduct a paired samples t-test the following syntax:

**ttest\_rel(a, b)**

where:

* a: an array of the sample observations from Group 1.
* b: an array of the sample observations from Group 2.

The test statistic is -2.9732 and the corresponding two-sided p-value is 0.0101.

Step 3: Interpret the results.

In this example, the paired samples t-test uses the following Null and Alternative Hypotheses:

H0: µ1 = µ2 (The mean pre-test and post-test scores are equal).

H1: µ1 ≠ µ2 (The mean pre-test and post-test scores are not equal).

Since the p-value (0.0101) is less than 0.05, we can reject the Null Hypothesis. Consequently, we have now sufficient evidence to say that the true scores are different for the students before and after participating in the study program.

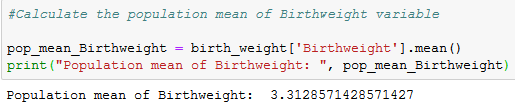
**Case Study**

We will be continuing the analysis of dataset labeled as “**birthweight reduced**”. Interestingly, we did not find anystrong and meaningful correlation between “Birthweight” and “mnocig” variables (Chapter I: Page 39). Such a conclusion cannot be based solely on the plot observations but has to be backed by the statistical tests.

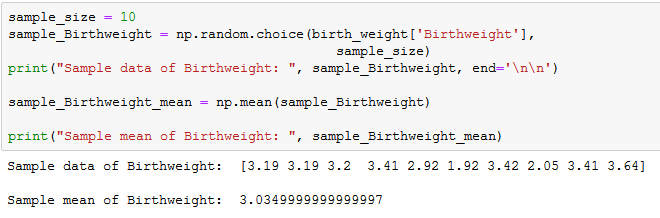
**Hypothesis Tests**

**Task 1:** **Estimating Average Birth Weight**

Step 1: Start with computing the average birth weight. Note that this value will serve in formulating the *null* hypothesis because, here you explicitly compute the population statistic-or the average birth weight. In most cases, such quantities are not directly observable and, in general, only the estimation for the population statistics is applied:



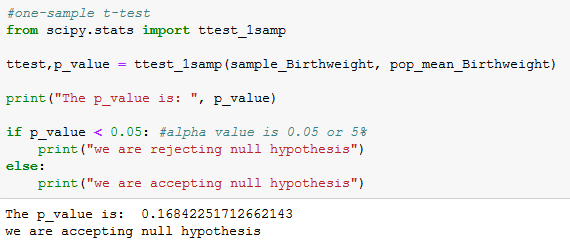
Step 2: We randomly sample the data and compute the sample mean.



Step 3: Now we perform the one sample t-test to check if there is any significant difference between the sample and population mean.

H0: No difference between population and sample mean.

H1: There is a significant difference between population and sample mean.



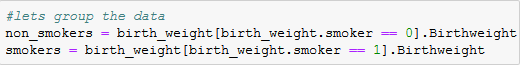
The computed p-value is equal to 0.168, which is much larger than the critical 0.05, and therefore, we cannot reject the null hypothesis.

**Task 2:** Compare the means of two independent groups.

Quite often, when performing statistical tests, we want to apply certain statistical tests on two different groups (for example, the average weight between men and women) and estimate whether there is a statistically-significant difference between the values obtained in the two groups. Let’s denote them with µ1 and µ2.

In this exercise, we will be testing the hypothesis on the “*Birthweight*” variable by grouping the data into smokers and non-smokers.

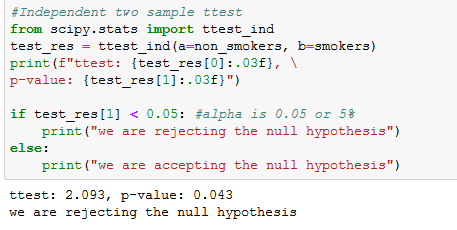
Step 1: We need to group the “*Birthweight*” data based on “smoker” (0: non-smoker, 1: smoker) variable.



Step 2: Perform the two sample t-test to compare the mean.

H0: Average birth weight of smoker’s group – Average birth weight of non-smoker’s group = 0

H1: Average birth weight of smoker’s group – Average birth weight of non-smoker’s group ≠ 0



The resulting p-value from this test is less than 0.05. As a conclusion, we can reject the *null* hypothesis and confirm that our initial observation (*i.e.,* Smoking status of the mother has no effect on the birth weight of the child) is wrong. There is a statistically significant difference between the birth weight of children whose mothers smoke, when compared to the children whose mothers do not smoke.

## Assessment

## Choose the appropriate option

1. **Following IQ scores are approximately normally distributed with a mean of 100 and standard deviation of 15. The proportion of people with IQs above 130 is:**
2. 95%
3. 68%
4. 5%
5. 2.5%
6. **Failing to reject the null hypothesis when it is false is:**
   1. Alpha
   2. Type I error
   3. Beta
   4. Type II error
7. **A statistic is:**
   1. A sample characteristic
   2. A population characteristic
   3. Unknown
   4. Normally distributed
8. **We reject the H0 Hypothesis when:**
   1. P-value < alpha
   2. P-value > alpha
   3. P-value = alpha
9. **People commonly lie when asked questions about personal hygiene.. This is an example of:**
   1. Sampling bias
   2. Confounding
   3. Non-response bias
   4. Response bias

## Fill in the spaces with appropriate answers

1. Because of the possibility of error in sampling from populations, researchers use \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ statements to describe results.
2. A t-test is used to compare \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.
3. The two forms of two sample t-tests are \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ and \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.
4. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ percent of the data will fall in two deviations in standard normal distribution.
5. An estimate of a population parameter given by two numbers between which the parameter would be expected to lie is called a \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ of the parameter.

## Programming Assignment

Using the data in the below URL,

<https://www.sheffield.ac.uk/polopoly_fs/1.937185!/file/Birthweight_reduced_kg_R.csv>

By referring to the code used in the case study perform the following tasks.

1. Find the correlation between “**Gestation**”and “**Birthweight**” variables.
2. Compute two sample t-test on the “**Gestation**” variable by grouping the data on “**smoker**”.

**Solutions:** Refer to page 69

## Solutions for Assessment

## Choose the appropriate options

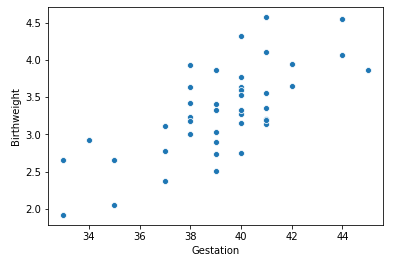
1. D
2. D
3. A
4. A
5. D

## Fill in the spaces with appropriate answers

1. Probability
2. Two means
3. Independent and dependent t-test
4. 95%
5. Interval Estimate

## Programming Assignment Solution

Task 1)



Task 2)



# CHAPTER 4: ANOVA

# Analysis of Variance

## Theory

In the previous chapter, we studied the continuous random variables, point estimate, confidence interval estimate, and test of hypothesis. Now, it is the time to get deeper into the analysis of variance.



## Assumptions for ANOVA

To utilize the ANOVA test, we have made the following assumptions:

* Sample for each group is drawn from a normally distributed population.
* All the populations have a common variance.
* All samples are drawn independently of each other.
* Within each sample, the observations are randomly collected and are independent of each other.
* Factor effects are additive.

## Why ANOVA?

To this point, we have compared two populations.

1. Independent samples t-test (random).
2. Matched sample t-test (paired).

What if we wish to compare the means of more than two populations? What if we wish to compare multiple populations with each one of them containing several subgroups?

Here comes ANOVA; **An**alysis **O**f **Va**riance, to answer all those questions.

Before we discuss the ANOVA, we need to understand the F-distribution.

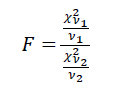


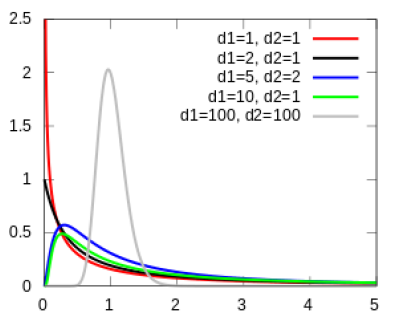
The F-distribution is a skewed distribution of probabilities similar to a chi-squared distribution. But whereas chi-squared distribution deals with the degree of freedom with one set of variables, the F-distribution deals with multiple levels of events having different degrees of freedom.

* 𝜒2 is useful in testing hypothesis about a single variance in a given population.
* Sometimes we want to test the hypothesis about the difference in variances of two populations: We ask,
  + Is the variance of 2 stocks the same?
  + Do parts manufactured in 2 shifts or on 2 different machines or in 2 batches have the same variance or not?
  + Is there a variability in the clinical response to the drug therapies of two samples?
* Ratio of 2 variance estimates:









## Hypothesis test for 2 sample variances

A machine produces metal sheets which are 22mm thick. There is a variability in thickness of different sheets due to machines, operators, manufacturing environment, raw materials, etc. The company wants to know the consistency of two machines and randomly sheets 10 sheets from machine 1 and 12 sheets from machine 2. The measurements for thickness are taken. Assume that the sheet thickness is normally distributed in the population.

The company wants to ascertain if the variance from each sample comes from the same population variance (population variances are equal) or from different population variances (population variances are unequal).

How do you test this hypothesis?

## Data

| **Machine 1** | | **Machine 2** | |
| --- | --- | --- | --- |
| 22.3 | 21.9 | 22.0 | 21.7 |
| 21.8 | 22.4 | 22.1 | 21.9 |
| 22.3 | 22.5 | 21.8 | 22.0 |
| 21.6 | 22.2 | 21.9 | 22.1 |
| 21.8 | 21.6 | 22.2 | 21.9 |
|  |  | 22.0 | 22.1 |

S12 = 0.11378, S22 = 0.02023



## What is a null and alternate hypothesis?



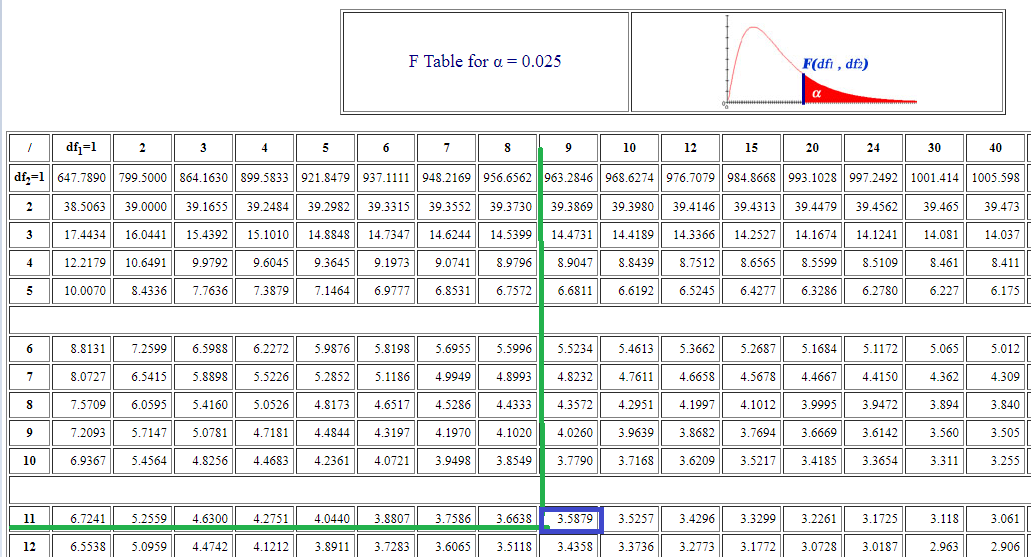
Is it a one-tailed test or a two-tailed test?

Two-tailed.

## What are numerator and denominator degrees of freedom?



## Reading an F-table

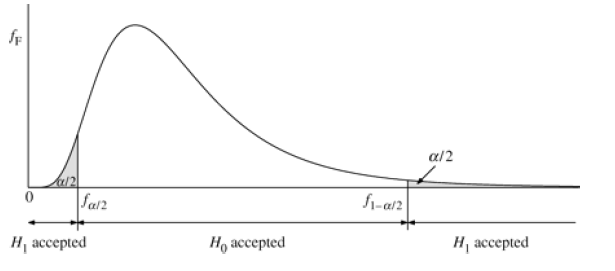


F-table: <http://socr.ucla.edu/Applets.dir/F_Table.html#:~:text=The%20F%20distribution%20is%20a,(12%2C10)%20).>









## Will you reject the null hypothesis or not?

Yes, we will reject the H0, because the population variances are not equal.

## What are the business implications?

Variance in machine 1 is higher than in machine 2. Machine 1 needs to be inspected for any technical issue which is causing variations.

Applications of F-Distribution:

* Test for equal variances.
* Test for differences of means in ANOVA.
* Test for regression models (Slopes relating one continuous variable to another, *e.g.,* Entrance exam scores and GPA).

Now let’s get back to ANOVA.

ANOVA uses the concept of “***Variance***” to compare the “***Means***” values.

## Comparing means

Suppose we want to compare 3 mean values to see if a potential difference exists among them.

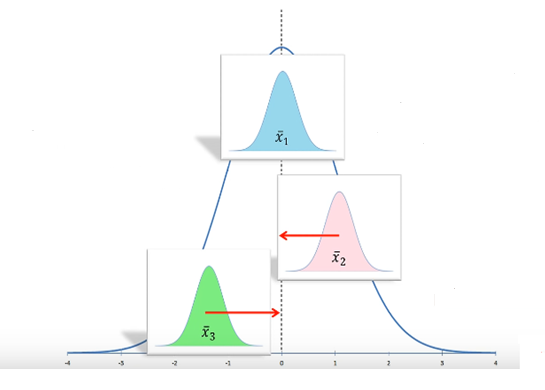
Is one mean value so far away from the other two that it is likely to be not from the same population?

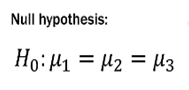


What we are asking is:

Do all the three mean values represent a common population? Or are all three so far apart from each other that they ALL likely come from unique populations?

## Comparing with population mean





Remember, we are not asking if the mean values are **EXACTLY** equal. We are asking if each mean value likely represents the larger overall population.

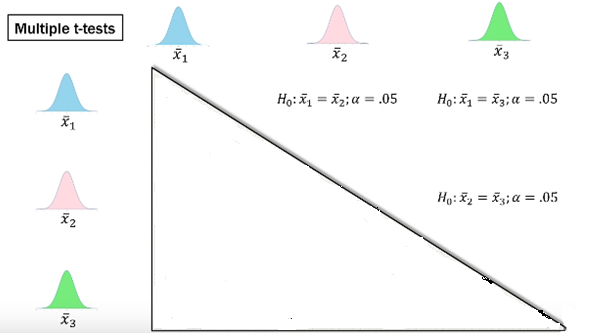
Variability **AMONG/BETWEEN** the sample means.

## Why can’t we compute the t-test three times?

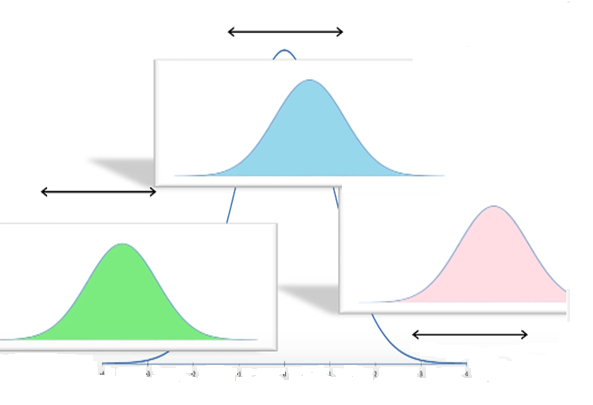
Pairwise comparison means three separate t-tests, and all of them have an α = 0.05. in such a case, the chances of a type I error is 95%.

But if we compute t-test three times, the errors of individual tests **COMPOUNDS** each time:

(0.95)(0.95)(0.95) = 0.857



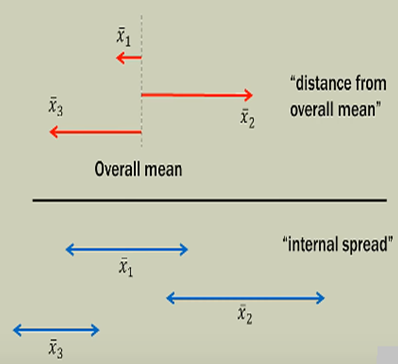
## What changed?



**VARIANCE or SPREAD** of each distribution.

Variability **AROUND/WITHIN** the distributions.

## ANOVA: Analysis of Variance is variability ratio.





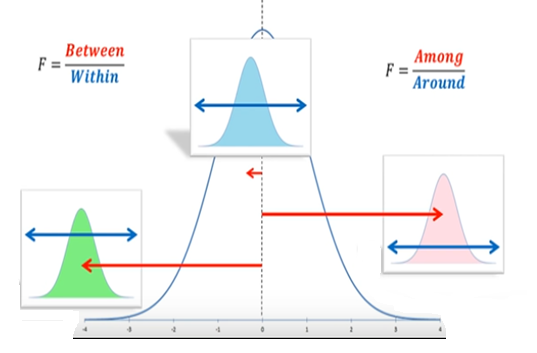
\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_



**Variance Between + Variance Within = Total Variance**

If the variability **BETWEEN** the mean values (distance from overall mean value) in the numerator is relatively large compared to the variance **WITHIN** the samples (internal spread) in the denominator, the ratio of the two will be much larger than 1. The samples then most likely do **NOT** come from a common population; So we can **REJECT THE NULL HYPOTHESIS** (or say that the mean values are equal).



Example: Let us say that 3 groups of students were given 3 different pills to enhance memories and their scores in a given exam were recorded. We want to understand if the differences in the scores are due to ‘’within’’ group differences or ‘’between’’ group differences.

|  | **Group 1** | **Group 2** | **Group 3** |
| --- | --- | --- | --- |
|  | 3 | 5 | 5 |
|  | 2 | 3 | 6 |
|  | 1 | 4 | 7 |
| **mean** | **2** | **4** | **6** |





Total Sum of Squares, SST

= (3−4)2+ (2−4)2+ (1−4)2+ (5−4)2+ (3−4)2+ (4−4)2+ (5−4)2+ (6−4)2+ (7−4)2 = 30

When there are “m” number of groups and “n” number of members in each group, the degrees of freedom are mn-1, which simply means that we can calculate one member by knowing the overall mean value.

How much of this variation is coming from within the groups and how much is coming from between the groups?

When there are “m” number of groups and “n” number of members in each group, the degrees of freedom are mn-1, since we can calculate one member knowing the group means.

Total Sum of Squares Between (SSB) = 3(2−4)2+3(4−4)2+3(6−4)2 = 24

When there are m groups, the degrees of freedom are **m – 1.**

**SST = SSW + SSB**

Also, for degrees of freedom, **mn-1 = m (n-1) + (m-1)**

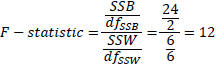
Given that the mean value of group 3 is the highest and that of group 1 the lowest, can we conclude that the pills given to group 3 had a larger impact or is it just the variation within the group accounting for the difference?

Let us have a null hypothesis that the population means values of the 3 groups from which the samples were taken are similar, i.e., the pills do not have an impact on the performance in the exam. µ1 = µ2 = µ3. Let us also have a significance level, α = 0.10.

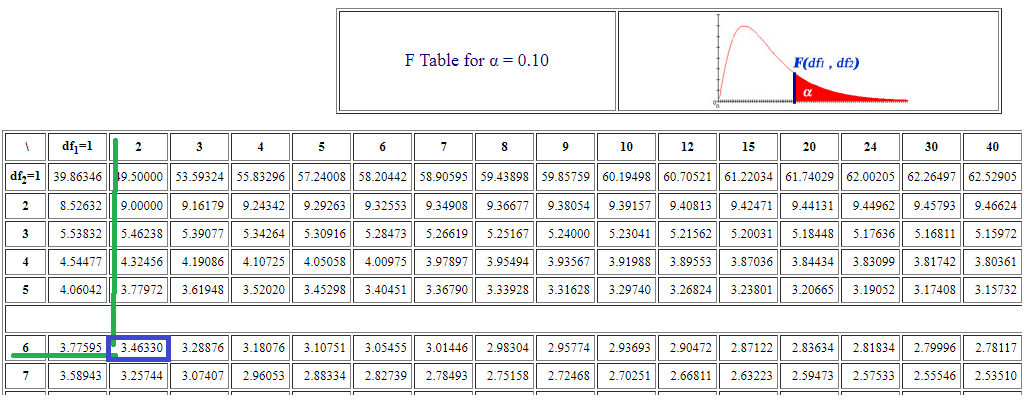
What is the alternate hypothesis?

The pills have a significant impact on performance.

The statistical test to test the alternate hypothesis is F-statistic.



If the numerator is much bigger than the denominator, it means that the variation between mean values has a bigger impact than the variation within the groups, thus rejecting the null hypothesis.



The df are 2 for numerator and 6 for denominator.

The critical F-statistic, therefore, is 3.46330. 12 is way higher than this and hence we can safely reject the null hypothesis. This means that the pills do have an impact on the performance of the students in the exam.



Tukey’s multiple comparison test is one of several tests that can be used to determine which mean value amongst a set of mean values differs from the rest. Tukey’s multiple comparison test is also called Tukey’s honestly significant difference test or Tukey’s HSD.

If the ANOVA leads to a conclusion that the evidence points towards a difference in the mean values of the groups, we might be interested in investigating which of the mean values are different. This is where the Tukey multiple comparison test is used. The test compares the difference between each pair of mean values with appropriate adjustment for the multiple testing. The results are presented as a matrix showing the result for each pair, along with P-values or as confidence intervals.

Tukey’s test is based on a formula that is very similar to the t-test. In fact, Tukey’s test is essentially a t-test, except that it corrects for family-wise error rate.

The formula for Tukey’s test is:



Where

* YA is the larger and the YB is the smaller of the two mean values being compared. And SE is the standard error of the sum of the means.
* This qs value can be compared to qvalue from the studentized range distribution. If the qs is obtained from the distribution, then the two mean values are said to be significantly different at level.



Since the *null* hypothesis for Tukey’s test states that all the mean values being compared are derived from the same population (i.e.  *μ*1 = *μ*2 = *μ*3 = ... = *μk*), the mean values should be normally distributed (according to the central limit theorem). This gives rise to the normality assumption of Turkey’s test.



We collect the data by blocking with other factors.

* In a one-way ANOVA, we selected a random sample from each column/treatment group.
* A Two-way ANOVA allows us to “*account for variations*” at the ROW level due to some other factor or grouping.
* By adding blocks or factors to the ROWS, we can “subtract out” the ROW variance from the overall ERROR variance.
* This subtraction allows us a greater factor on COLUMN or GROUP differences, which makes it easier to detect group differences.
* We are attempting to minimize the ERROR variance by saying, “*Hey now, some of the ERROR variance is actually due to the variance in the ROWS*!”
* So we will now have 4 types of Sum of Squares / SOURCES OF VARIANCE
  1. Sum of the Squares Total.
  2. Columns/Groups.
  3. Rows/Blocks.
  4. Error.

Formulation of two-way ANOVA (without replication) is as follows,

|  | **Sum of Squares** | **Degrees of Freedom** | **Mean Sum of Squares** |
| --- | --- | --- | --- |
| **SSC** | **Sum of Squares (Columns/treatments)** | **df (Columns) = C - 1** | **MSC = (SSC/df columns)** |
|
| **SSB** | **Sum of Squares (Blocks)** | **df (Blocks) = B - 1** | **MSB = (SSB/df Blocks)** |
|
| **SSE** | **Sum of Square (Within/error)** | **df (error) = (C-1)(B-1)** | **MSE = (SSE/df error)** |
|
| **SST** | **Sum of Squares (Total)** | **df (total) = N - 1** | **F = MSC/MSE** |
|

Where

* N = total number of observations.
* C = No of Columns/Treatments.
* B = No of Blocks.

Example:

| **Production of Three varieties of wheat (MT/Acre)** | | | |
| --- | --- | --- | --- |
| **Fertilizers** | **Seeds** | | |
| **A** | **B** | **C** |
| **Agri Gold** | **6** | **5** | **5** |
| **Surya** | **7** | **5** | **4** |
| **Arthi** | **3** | **3** | **3** |
| **Deepak** | **8** | **7** | **4** |

* Outcome (result): Production of three varieties of wheat
* Factors affecting the results are two :
  1. Variety of Seeds (Wheat).
  2. Variety of Fertilizers.
* Null Hypothesis: Production of all three varieties are equal in proportion.
* Significance level = 0.05.

| **Production of Three varieties of wheat (MT/Acre)** | | | | |  |
| --- | --- | --- | --- | --- | --- |
| **Fertilizers** | **Seeds** | | | |  |
|  | | **A** | **B** | **C** | **Row Sum** |
| **Agri Gold** | | **6** | **5** | **5** | **16** |
| **Surya** | | **7** | **5** | **4** | **16** |
| **Arthi** | | **3** | **3** | **3** | **9** |
| **Deepak** | | **8** | **7** | **4** | **19** |
| **Column Sum** | | **24** | **20** | **16** |  |

Step 1: *Null* and Alternative hypothesis

**H0: µ1 = µ2 = µ3**

(Variety of seed and fertilizer has no significant effect on the production of wheat)

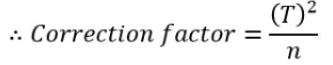
**H1: Not all three are equal**

(Variety of seed and fertilizer has a significant effect on the production of wheat)

Step 2: Sum of Squares of Total Variance

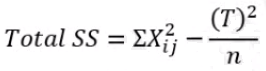
n = 12, T = 6+7+3+8+5+5+3+7+5+4+3+4 = 60

Here we use the correction factor to calculate the sum of squares of total variance instead of mean value method:



= (60)2/12

**Correction factor = 300**

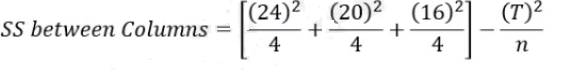


Total SS = (36+49+9+64+25+25+9+49+25+16+9+16) – 300

= 332 – 300

**Total SS = 32**

Step 3: Sum of the squares between columns

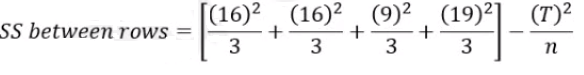


= [144 + 100 + 64] – 300

=308 – 300

**SS between columns = 8**

Step 4: Sum of the squares between Rows



= [85.33 + 85.33 + 27 + 120.33] – 300

= 318 – 300

**SS between rows = 18**

Step 5: Sum of the squares Residual or Error

SS residual or error = Total SS – (SS between columns + SS between rows)

= 32 – (8 + 18)

**SS residual or error = 6**

|  | **Sum of Squares** | **Degrees of Freedom** | **Mean Sum of Squares** | **F ratio** | **F tab** | **H0** |
| --- | --- | --- | --- | --- | --- | --- |
| **SSC** | **8** | **C - 1 = 3 -1 = 2** | **MSC = 8/2 = 4** | **MSC/MSE = 4/1 = 4** | **F(2,6) = 5.14** | **can't be rejected** |
|
| **SSB** | **18** | **B - 1 = 4 - 1 = 3** | **MSB = 18/3 = 6** | **MSB/MSE = 6/1 = 6** | **F(3,6) = 4.76** | **rejected** |
|
| **SSE** | **6** | **(C-1)(B-1) = (2)(3) = 6** | **MSE = 6/6 = 1** |  |  |  |
|  |  |  |
| **SST** | **32** | **N - 1 = 12 -1 = 11** |  |  |  |  |
|  |  |  |

Conclusion:

* Variety of fertilizers could significantly affect the productions of A, B, & C.
* Variety of seeds could not significantly affect the productions of A, B, & C.



Example: The research was conducted for the reduction in blood pressure (BP) (in mm of mercury) by three drugs X, Y, and Z on three groups of people A, B, and C.

The experiment is repeated twice for each drug.

| Amount of Blood Pressure  Reduction (mm) of mercury | | | |
| --- | --- | --- | --- |
|
| Group of people | Drugs | | |
| X | Y | Z |
| A | 14 | 10 | 11 |
| 15 | 9 | 11 |
| B | 12 | 7 | 10 |
| 11 | 8 | 11 |
| C | 10 | 11 | 8 |
| 11 | 11 | 7 |

* Outcome (result): Effectiveness of three drugs (X, Y, Z) in reducing blood pressure (BP).
* Number of factors affecting the results: 2
  1. Different types of drugs
  2. Drugs are tested on different groups of people
* Null hypothesis: BP is reduced by the drugs X, Y, and Z to the same levels.
* Significance level: 5%

Find the answers

1. Do the drugs act differently?
2. Are the different groups of people affected differently?
3. Is the interaction term significant?

**Solutions:**

**Null and Alternative hypothesis**

H0: The reduction in BP by the three drugs is equal

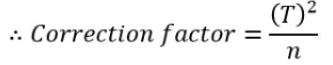
(There is no effect of two factors like; types of drugs and groups of people in reducing BP)

H1: The reduction in BP by three drugs is different

(There is a significant effect of two factors like; types of drugs and groups of people on the reduction in BP)

Step 1: Correction Factor

n = 18, T = 187



**Correction factor = (187)2/18 = 1942.72**

Step 2: Total Sum of Squares

Total SS = [(14)2+ (15)2 + (12)2 + (11)2 + (10)2 + (11)2 + (10)2 + (9)2 + (7)2 + (8)2 +

(11)2 + (11)2 + (11)2 + (11)2 + (10)2 + (11)2 + (8)2 + (7)2] – [(187)2/18]

= [2019] – [1942.72]

**Total SS = 76.28**

| Amount of Blood Pressure  Reduction (mm) of mercury | | | |  |
| --- | --- | --- | --- | --- |
|  |
| Group of people | Drugs | | |  |
| X | Y | Z | Total |
| A | 14 | 10 | 11 | 70 |
| 15 | 9 | 11 |
| B | 12 | 7 | 10 | 59 |
| 11 | 8 | 11 |
| C | 10 | 11 | 8 | 58 |
| 11 | 11 | 7 |
| Total | 73 | 56 | 58 |  |

Step 3: Sum of the Squares between columns (i.e. between drugs)

= [(73)2/6 + (56)2/6 + (58)2/6] - [(187)2/18]

= [888.16 + 522.66+ 560.67] – [1942.72]

= 1971.19 – 1942.72

**Sum of Squares between columns = 28.77**

Step 4: Sum of the Squares between rows (i.e. between people)

= [(70)2/6 + (59)2/6 + (58)2/6] - [(187)2/18]

= [816.67 + 580.16 + 560.67] – [1942.72]

= 1957.50 – 1942.72

**Sum of Squares between rows = 14.78**

Step 5: Sum of the Squares within samples

= [(14 – 14.5)2 + (14 – 14.5)2 + (10 – 9.5)2 +(9 – 9.5)2 + (11 – 11)2 +

(11 – 11)2 + (12 – 11.5)2 + (11 – 11.5)2 + (7 – 7.5)2 + (8 – 7.5)2 +

(10 – 10.5)2 + (11 – 10.5)2 + (10 – 10.5)2 + (11 – 10.5)2 + (11 – 11)2 +

(11 – 11)2 + (8 – 7.5)2 + (7 – 7.5)2]

= 0.25 + 0.25 + 0.25 + 0.25 + 0 + 0 + 0.25 + 0.25 + 0.25 + 0.25

+ 0.25 + 0.25 + 0.25 + 0.25 + 0 + 0 + 0.25 + 0.25

**Sum of the Squares within Samples = 3.5**

Step 6: Sum of the Squares for Interaction Variation.

**Interaction** effects occur when the effect of one variable depends on the value of another variable. **Interaction** effects indicate that a third non-related variable influences the relationship between an independent and a dependent variable.

= 76.28 – [28.77 + 14.78 + 3.50]

= 76.28 - 47.05

**Sum of the Squares for Interaction variation = 29.23**

H0: The reduction in BP by the three drugs is equal.

H1: The reduction inBPby the three drugs is different.

* SS Total = 76.28
* SS between columns = 28.77
* SS between rows = 14.78
* SS within samples = 3.50
* SS for interaction variation = 29.33
* N = 18
* Number of columns (c) = 3
* Number of rows (r) = 3

|  | **Sum of Squares** | **Degrees of Freedom** | **Mean Sum of Squares** | **F ratio** | **F tab** | **H0** |
| --- | --- | --- | --- | --- | --- | --- |
| **SSC** | **28.47** | **C - 1 = 3 -1 = 2** | **MSC = 28.77/2 = 14.385** | **MSC/MSE = 14.385/0.389 = 36.9** | **F(2,9) = 4.26** | **rejected** |
|
| **SSB** | **14.78** | **B - 1 = 3 - 1 = 2** | **MSB = 14.78/2 = 7.390** | **MSB/MSE =  7.390/0.389 = 19.0** | **F(2,9) = 4.26** | **rejected** |
|
| **Interaction** | **29.23** | **2 x 2 = 4** | **29.23/4 = 7.308** | **7.308/0.389 = 18.79** | **F(4,9) = 3.63** | **rejected** |
|
| **SSE** | **3.5** | **(18 -9) = 9** | **3.50/9 = 0.389** |  |  |  |
|  |  |  |

Conclusion:

* All three F-ratios are significant (p < 0.05) for all three cases.
* Hence Null hypothesis is rejected in all three cases and it implies that:
  1. The drugs act differently (they reduce BP to different degrees).
  2. Different groups of people are affected differently (The reduction in BP is different for different groups of people).
  3. Interaction term is significant.

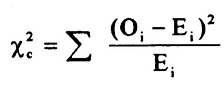
Always look for the significant interaction First. If an interaction exists and is significant, the effects of row and column cannot be evaluated individually.



Chi-square uses a test statistic to compare the difference between the expected and the actual, and then returns a probability of getting observed frequencies. It helps us to understand the relationship between the two categorical variables; grade level, gender, age group, *etc*.

We use the chi-square distribution and critical value to accept or reject the null hypothesis.

Formula



Where

* The subscript ‘c’ is the degrees of freedom.
* ‘O’ is the observed value.
* ‘E’ is the expected value.

Example: Data of experiment based on vaccination against a viral infection:

|  | Infected | Not Infected |
| --- | --- | --- |
| Vaccinated | 31 | 469 |
| Not vaccinated | 185 | 1315 |

Test the effectiveness of vaccination in preventing the spread of the infection at significance of 5%.

## Solutions:

## Null and Alternative hypothesis

H0: Vaccination is not effective against the infection.

(Vaccination and infection are independent or the vaccine is not effective to prevent the infection to spread)

H1: Vaccination is effective against the infection.

(Vaccination and infection are associated; vaccine is effective to prevent the infections to spared)

## Observed Frequencies

| Observed frequencies | Infected | Not Infected | Total |
| --- | --- | --- | --- |
| Vaccinated | 31 | 469 | 500 |
| Not vaccinated | 185 | 1315 | 1500 |
| Total | 216 | 1784 | N = 2000 |

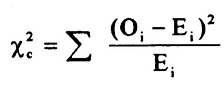
## Expected Frequencies

| Expected frequencies | Infected | Not infected | Total |
| --- | --- | --- | --- |
|
| Vaccinated | (500x216)/2000 = 54 | (500x1784)/2000 = 446 | 500 |
|
| Not Vaccinated | (1500x216)/2000 = 162 | (1500x1784)/2000 = 1338 | 1500 |
|
| Total | 216 | 1784 | N = 2000 |
|

## Chi-square

| **Cell No.** | **Observed frequency** | **Expected frequency** | **(Oij - Eij)** | **(Oij - Eij)2** | **(Oij - Eij)2/Eij** |
| --- | --- | --- | --- | --- | --- |
| **(Oij)** | **(Eij)** |
| 1 | 31 | 54 | 31 - 54 = -23 | 529 | 529/54 = 9.796 |
|
| 2 | 469 | 446 | 469 - 446 = 23 | 529 | 529/446 = 1.186 |
|
| 3 | 185 | 162 | 185 - 162 = 23 | 529 | 529/162 = 3.265 |
|
| 4 | 1315 | 1338 | 1315 - 1338 = -23 | 529 | 529/1338 = 0.395 |
|

**Conclusion:**

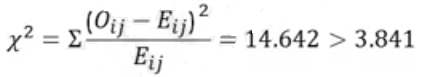


= (9.796+1.186+3.265+0.395)

**ᶍ2= 14.642**

* Significance level 5% or α = 0.05
* Degrees of freedom: (r – 1)x(c – 1) = (2 - 1) X (2 – 1) = 1

**ᶍ2Crit = 14.642**



H0: Vaccination is not effective. Rejected

H1: Vaccination is effective. Accepted

**The vaccine is effective against the spread of the infection**

**AIM**

The aim of the following lab exercise is to write python code so that we can get hands on descriptive statistics.

The labs for this chapter include the following exercises:

* One-way ANOVA.
* Two-way ANOVA.
* *Chi-square* test of independence.

We will be working with python3 and jupyter notebook IDE.

## Case Study

The data was extracted from the 1974 *Motor Trend* US magazine, and contains details about fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models). We are going to analyze the effects of independent variables such as the “Number of Cylinders”, “Engine”, and “Transmission” on the dependent variable “MPG” Miles/gallon.

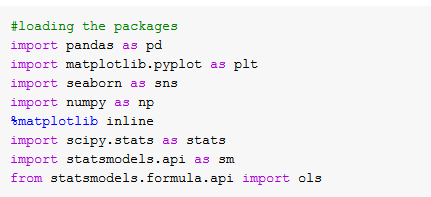
You can download the data from the below link:

<https://www.kaggle.com/ruiromanini/mtcars>

General description of the data

| **#** | **Var** | **Description** | **Comments** |
| --- | --- | --- | --- |
| 0 | name | Model of Vehicle | The car types are a mix that includes sedans (Datsun, Ford, Honda,…), luxury sedans (Mercedes, Cadillac,..), muscle cars (Javelin, Challenger, Camero…), and high-end sports cars (Porsche, Lotus, Maserati, Ferrari…) |
| 1 | mpg | Miles/US Gallon | mpg is the determinant of fuel efficiency |
| 2 | cyl | Number of cylinders | Data includes vehicles with 4,6,8-cylinder engines. |
| 3 | disp | Displacement (cu.in.) | Displacement measures the overall volume in the engine as a factor of cylinder circumference, depth, and total number of cylinders. This metric gives a good proxy for the total amount of power the engine can generate. |
| 4 | hp | Gross horsepower | Gross horsepower measures the theoretical output of an engine’s power output; notably, *gross* rating is of the engine in an isolated environment outside any specific vehicle. When installed in a car, exhaust systems, carburetor, alternator, power systems, etc. all influence the power that actually gets to the drive train. Moreover, according to online sources, in the early 1970s, regulatory changes influenced how gross horsepower was measured. As this dataset is from the early-mid 1970s, it’s unclear if hp metrics may be used as reliable comparators of the engine power across models as it is uncertain how the manufacturers are reporting. |
| 5 | drat | Rear-axle ratio | The rear axle gear ratio indicates the number of turns of the drive shaft for every one rotation of the wheel axle. A vehicle with a high ratio would provide more torque and thus more towing capability, for example. Transmission configuration can often influence a manufacturer’s gearing ratio. |
| 6 | wt | Weight (lb/1000) | The overall weight of the vehicle per 1000lbs (half US ton) |
| 7 | qsec | 1/4 mile time | A performance measure, primarily of the acceleration. Fastest time to travel 1/4 mile from standstill (in seconds). |
| 8 | vs | V/S | Binary variable signaling the engine-cylinder configuration a V-shape (vs=0) or Straight Line (vs=1). V==0 and S==1. Configuration offers tradeoffs in a power/torque, design usage in terms of space/size of engine, and performance or center of gravity of vehicle. The geometry and placement of the engine, as influenced by its cylinder head, can have numerous knock-on influences on the vehicle beyond the consideration of the technical engineering of the cylinder angle. |
| 9 | am | Transmission Type | A binary variable signaling whether a vehicle has automatic (am=0) or manual (am=1) transmission configuration. |
| 10 | gear | Number of forward gears | Number of gears in the transmission. Manual transmissions have either 4 or 5 forward gears, while the automatic transmission either has 3 or 4. |
| 11 | carb | Number of carburetors | The number of carburetor barrels. Engines with higher displacement typically have higher barrel configuration to accommodate the increased airflow rate of the larger engine. In other words, more capacity is available for an engine when it may need the constraining power output with limited barrels. A vehicle may have multiple physical carburetors, but such design is less common; this metric is the sum of the number of carburetors and the number of barrels inside the carburetors. |

Step 1: loading the necessary packages.

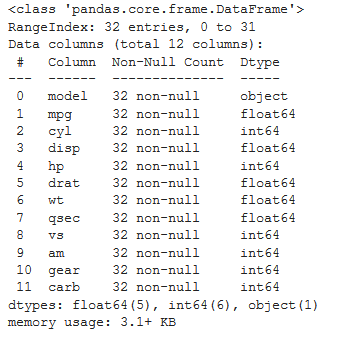


Step 2: Importing the data.



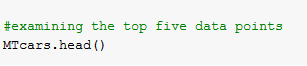
Step 3: Exploring the dataset using info() function.

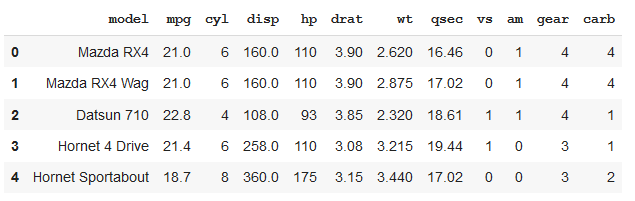




We have 32 observations, 12 variables, and 0 null variables.

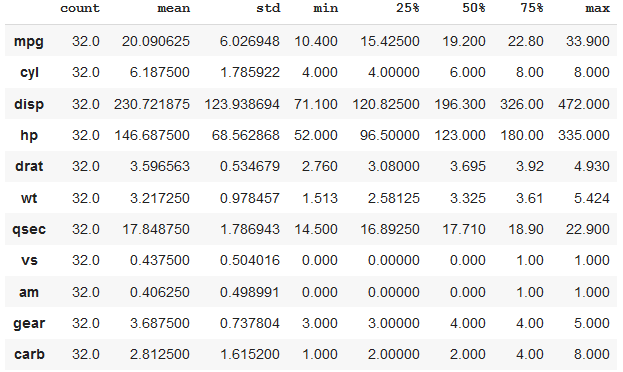
Step 4: Let’s explore the top five observations.



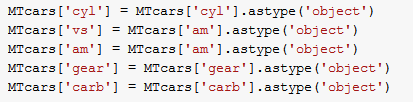


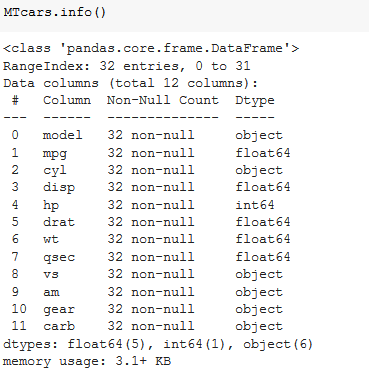
Step 5: Computing the descriptive statistics





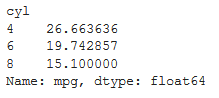
Step 6: We need to convert the data types of cyl, vs, am, gear, and carb variables from ‘int64’ to ‘object’, so that we can group them together and perform the analysis.





Step 7: Now, let’s group the data based on the ‘cyl’ (Number of cylinders) variable and compute the mean values of ‘mpg’ for each level.





The mean mpg value (Miles/US Gallon) for vehicles with 4 cylinders is 26.66, for 6 cylinders is 19.74, and for 8 cylinders is 15.10.

Step 8: Visualizing the distribution of mpg variable by grouping on cyl.





The average mpg value for each group of cylinders is different. Let’s compute ANOVA to check whether the difference is statistically significant or not.

Step 9: fit the linear regression model between mpg and cyl.



We have fitted the linear regression model between ols (ordinary least square) function.

Step 10: Framing the Hypothesis.

**H0: µ4 = µ6 = µ8**

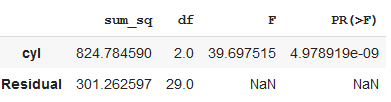
(Variety of Cylinders has no significant effect on the mpg (Miles/US Gallon))

**H1: Not all three are equal**

(Variety of Cylinders has a significant effect on the mpg (Miles/US Gallon))

Step 11: Let’ perform a one way ANOVA.





Using the “anova\_lm” function we have computed one way ANOVA. The p-value is less than 0.05.

Conclusion:

We can reject the null hypothesis. The variety of cylinders has a significant effect on the mpg.

## **Task 2:** **Two way ANOVA**

Let’s introduce a block variable ‘am’ to account for the variation at the ROW level.

Step 1: Fit the linear regression model between mpg, cyl, and am.



Step 2: Framing the Hypothesis.

**H0: µ4 = µ6 = µ8**

H0: **µ0 = µ1**

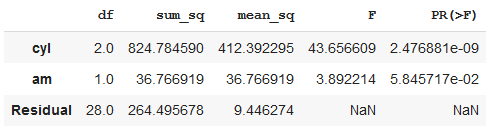
(The variety of Cylinders and Transmission Type has no significant effect on the mpg (Miles/US Gallon))

**H1: Not all three are equal**

(The variety of Cylinders and Transmission Type has a significant effect on the mpg (Miles/US Gallon))

Step 3: Let’s perform the Two-Way ANOVA without replication.





The p-values for Cylinders and Transmission Type variables are less than 0.05.

Conclusion:

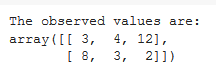
We can reject the null hypothesis. The variety of Cylinders and Transmission Type has a significant effect on the mpg (Miles/US Gallon).

## Task 3: Chi-square test of independence

Perform the chi-square test of independence to check whether the two categorical variables, the Cylinders (cyl) and Transmission Types (am) have any association between them.

Step 1: Compute the cross table between Cylinder and Transmission Types.



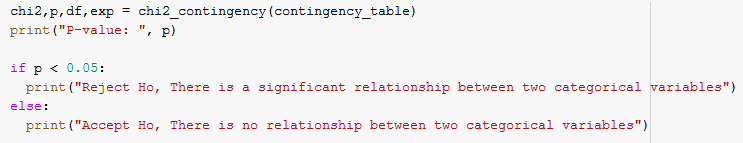


Step 2: Framing the hypothesis

H0: There is no relationship between the two categorical variables.

H1: There is a significant relationship between the two categorical variables.

Step 3: Perform the chi-square test of independence.





Conclusion:

We can reject the H0. Thus, there is a significant association between “cyl” (Cylinders) and “am” (Transmission Type).

## Assessment

## Choose the most appropriate option

1. **Analysis of variance is a statistical method of comparing several populations.**
2. Means
3. Variances
4. Standard Deviations
5. None of The Above
6. **In a study, subjects are randomly assigned to one of the following three groups: control, experimental A, or experimental B. After treatment, the mean scores for the three groups are compared. The appropriate statistical test for comparing these means scores is:**
   1. The Analysis of Variance
   2. The Correlation Coefficient
   3. Chi Square
   4. The t-test
7. **Assuming that the null hypothesis being tested by ANOVA is false, the probability of obtaining an F-ratio that exceeds the value reported in the F table as the 95th percentile is:**
   1. Equal to .05
   2. Greater than .05
   3. Less than .05
   4. None of these
8. **The error deviations within the SSE statistic measure distances:**
   1. Between groups
   2. Within groups
   3. Both (a) and (b)
   4. None of the above
9. **As the variability due to chance decreases, the value of F will**
   1. Decrease
   2. Stay the same
   3. Increase
   4. Can’t Tell From The Given Information

## Fill in the spaces with appropriate answers

1. While a t-test is used to compare two mean values, the one-way ANOVA can be used to simultaneously compare \_\_\_\_\_\_\_\_\_\_\_ groups.
2. To find the MSC, we divide the SSC by \_\_\_\_\_\_\_\_\_.
3. The \_\_\_\_\_\_\_\_\_\_\_\_ sum of squares measures the variability of the observed values around their respective treatment mean values.
4. The \_\_\_\_\_\_\_\_\_\_\_\_ sum of squares measures the variability of the sample treatment means around the overall mean value.
5. In Two way ANOVA, SST = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ + \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ + Error.

## Programming Assignment

Using the data in the below URL.

<https://www.kaggle.com/ruiromanini/mtcars>

Task 1: Compute a one way ANOA between mpg and gear.

Task 2: Compute the two way ANOVA between, mpg, gear, and vs.

Task 3: Compute the chi-square test between gear and vs variables.

**Solutions:** Refer to page 181

## Solutions for Assessment

## Choose the appropriate options

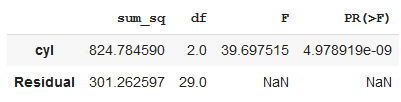
1. A
2. A
3. C
4. B
5. C

## Fill in the spaces with appropriate answers

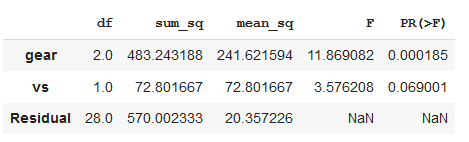
1. More than two
2. C – 1 degrees of freedom
3. Error
4. Treatment
5. SS columns, SS rows

## Programming Assignment Solution

Task 1)



Task 2)



Task 3)



# CHAPTER 5: Data Cleaning and Exploratory Data Analysis

## Theory

In the previous chapter, we learned the analysis of variance. In this chapter, we will be working with data cleaning and the exploration to gain some meaningful insights into the data.



Contrary to the common perception of the data we have handled in this course so far, , not every dataset is a perfectly curated group of observations with no missing values or anomalies. Real-world data is messy and requires cleaning and wrangling it into an acceptable format before we start the analysis. Data cleaning is an un-glamorous, but necessary part of most actual data science problems.

We will understand this chapter by going through a case study.

## Case Study

## Problem Definition

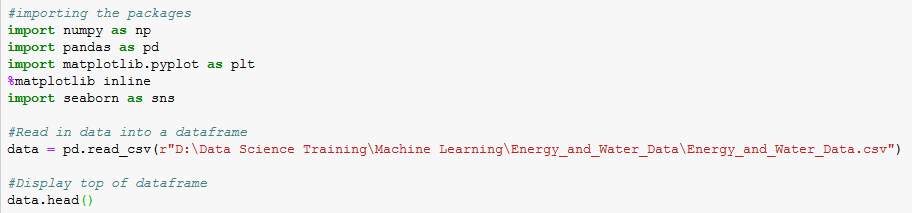
The first step before we obtain coding is to understand the problem that we are trying to solve with the available data. In this case study, we will work with publicly available **Building Energy** data from New York City.

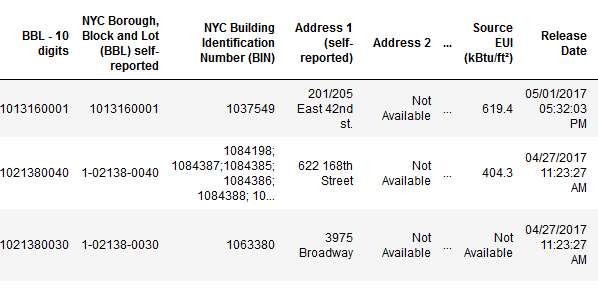
**Data:** <https://www1.nyc.gov/html/gbee/downloads/misc/nyc_benchmarking_disclosure_data_reported_in_2017.xlsb>

General Description of the data.

<http://www.nyc.gov/html/gbee/downloads/misc/nyc_benchmarking_disclosure_data_definitions_2017.pdf>

First, we can load the data as Pandas Dataframe and take a look:





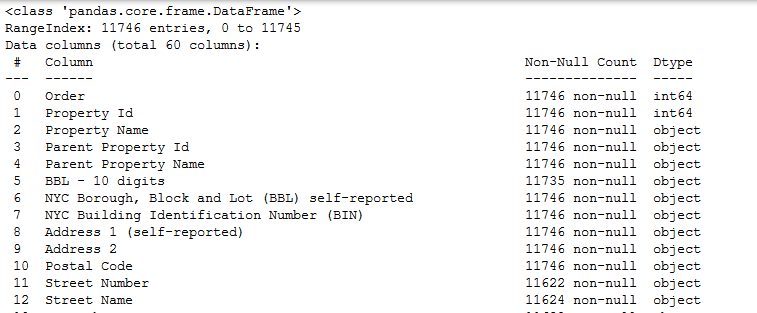
This is a subset of the full data which contains 60 columns. We can already see a couple of issues here: first, we know that our aim is to predict the **Energy Star score** but we don’t have the mean values for any of the columns. While this isn’t necessarily an issue --- we can get around this problem by making an accurate model of the data without any knowledge of the variables. This has practical implications since at the end of the day, we want to focus on interpretability, and it might be important to understand at least some of the columns.

We should at least understand the **Energy Star score**, which is described as:

A 1-to-100 percentile ranking based on the self-reported energy usage for the reporting year. The **Energy Star score** is a relative measure employed for comparing the energy efficiency of the buildings.

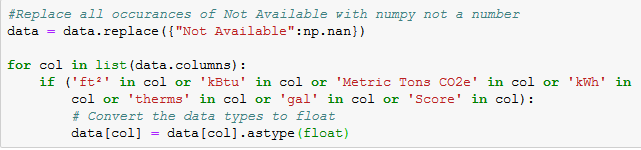
That clears up the first problem about data. However, we have another issue about the missing values in the data, which are encoded as “Not Available”. This is a “string” in Python which means that even the columns with numbers will be stored as “object” data types because Pandas converts a column with any string into a column of all strings. We can obtain the data types of the columns using the following command





As expected, some of the columns that clearly contain numbers (such as ft2), are stored as objects. We can’t do numerical analysis on the strings, so these columns need to be converted to number (specially float) data types before any analysis can be performed!

Here is a little Python code that replaces all the “*Not Available*” entries with ‘*not a number’* (np.nan), which can be interpreted as numbers, and then convert the relevant columns to the *float* data type.

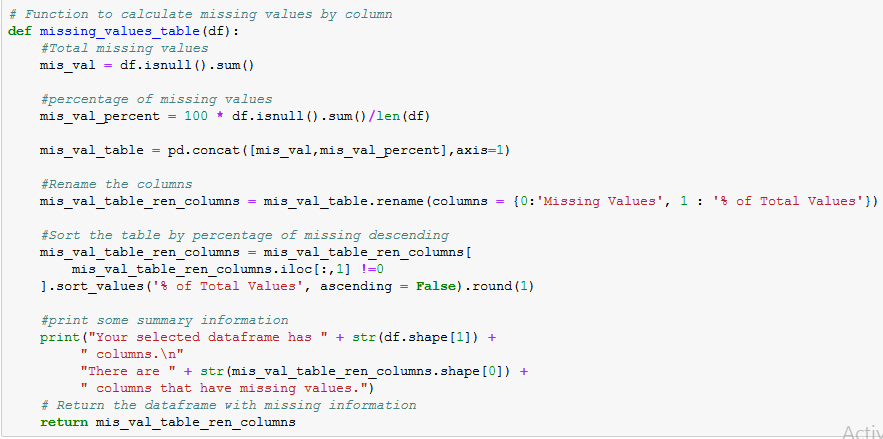


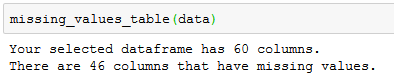
Once the correct columns have numbers, we can start to statistically analyze the data.

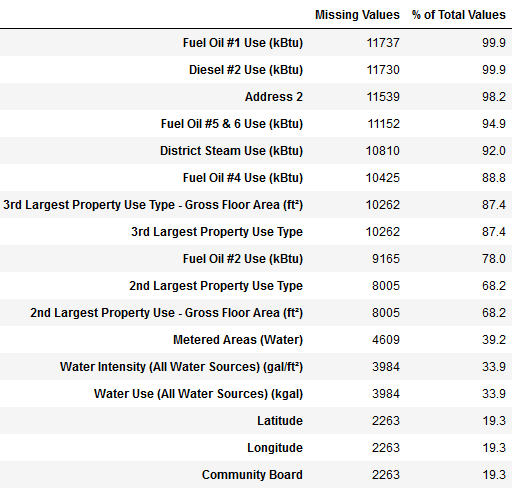


In addition to incorrect data types, missing values is another common problem when dealing with real-world data. These can arise for many reasons and have to be either filled in or removed before we train a machine learning model to analyse the data.

First, let’s get a sense of how many missing values are there in each column.

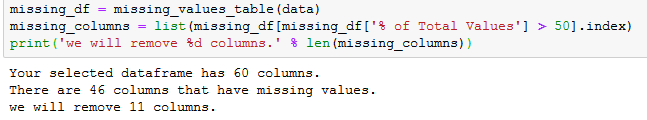






While we always have to be careful when removing information, if a column has a high percentage of missing values, there is a possibility that it might not be useful to our model.

After removing the missing value, the remaining data in a column may be slightly arbitrary. So for this case study; we will remove any columns which have more than 50% of values as missing. In general, be careful about dropping any information because even if it is not there for all the observations, it may still be useful for predicting the target value.





## Imputing Missing Values

While we dropped the columns with more than 50% missing while cleaning data, we still have quite a few missing observations. Machine learning models cannot deal with any absent values, so we have to fill them in, a process known as **imputation.**

Every value that is *Nan*, represents a missing observation. While there are a number of ways to fill in missing data, we will use a relatively simple method, called median imputation for numeric variables and most frequent (mode) imputation for categorical variables. This method replaces all the missing values in the columns with the median and mode values.

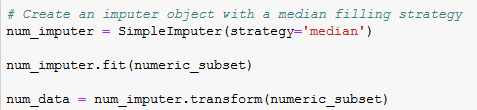
To do this, first, we need to subset the categorical and numerical variables for the imputation process.

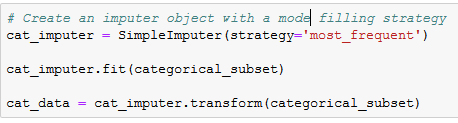


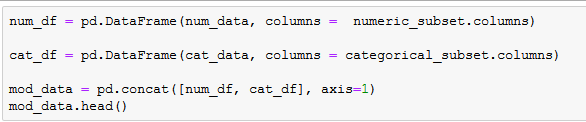
In the following code, we create a *Scikit-Learn* Imputer object with the strategy set to “median” (numeric variable) and “*most\_frequent*” (categorical variables) number. Subsequently, we then train this object on the numeric and category data (using *imputer.fit*) and use it to fill in the missing values in both the numeric and category data (using *imputer.transform*). This means that the missing values in the numeric and category data are filled with the corresponding median and mode values.

Import the *Scikit-Learn* package before we start the imputation process.









We concatenated the data and saved it as a modified data set.

Now let’s check the presence of any null values in the data.



We can see that there are no null values present in the data set.

**Note:** Theoretically, up to 25 to 30% is the maximum number of missing values in a data set. If the number of missing values is higher, we might want to drop the variable from analysis. However, this rule varies in practice. At times we have a dataset with ~50% of missing values but still, the customer insists on using it for analysis. In such cases for practical purposes, we treat the dataset on a case to case basis.



At this point, we may also want to remove outliers. However, we have to be careful not to throw away measurements just because they look strange. This is important because the measurements may be the result of actual phenomena that require further investigation. When removing outliers, I try to be as conservative as possible, using the definition of an extreme outlier:

On the low end, an extreme outlier is below **Q1 – 3 \* IQR**

On the high end, an extreme outlier is above **Q3 + 3 \* IQR**

An outlier in a dataset can also be due to typos in data entry, mistakes in units, or in rare cases they could be legitimate but extreme values.

Now that the tedious --- but necessary --- step of data cleaning is complete, we can move on to explore our data!

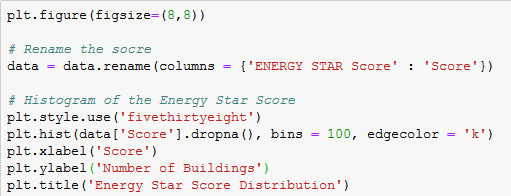


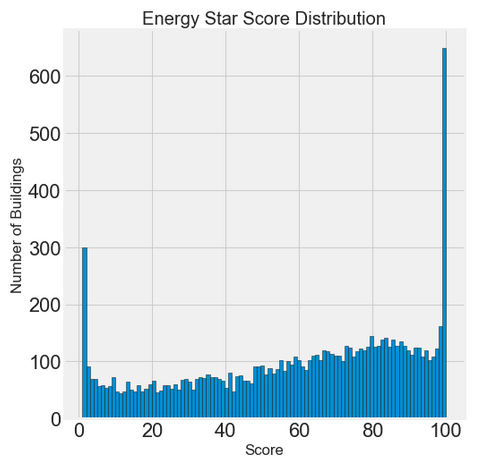
In short, the goal of EDA is to learn what our data can tell us. Usually, the EDA starts with a general overview, which is then gradually narrowed to specific areas as we find interesting trends and parts in the data. The findings may be interesting in their own right, or they can be used to select our modeling choices, such as by helping us to decide which features to use for further analysis.

## Univariate analysis (Single variable plots)

A single variable (called univariate) plot shows the distribution of a single variable such as in a histogram.

The goal is to predict the Energy Star score (Renamed to score in our data), so a reasonable place to start the analysis is to examine the distribution of this variable. A histogram is a simple yet effective way to visualize the distribution of a single variable and is easy to make using *matplotlib*.



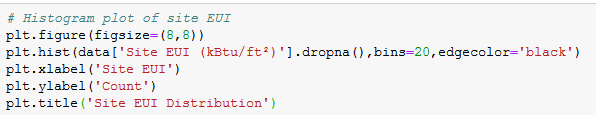


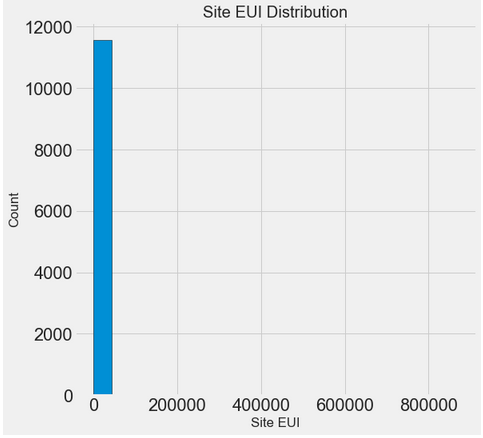
This graph looks quite suspicious! The Energy Star score is a percentile rank, which means that we would expect to observe a uniform distribution, with each score assigned to the same number of buildings. However, a disproportionate number of buildings have either the highest, 100, or the lowest, 1, score (higher is better for the Energy Star score).

If we look back at the definition of the score, we observe that it is based on “*self-reported energy usage*” which might explain the very high scores. Asking the building owners to report their own energy usage is like asking the students to report their own scores on a test. Consequently, this probably is not the most objective measure of a building’s energy efficiency.

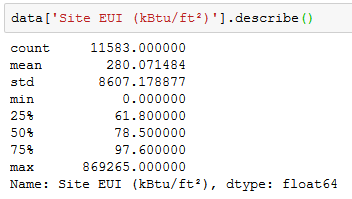
If we had an unlimited amount of time, we might want to investigate the reasons behind very high and very low scores in several buildings in our dataset. However, our objective is only to predict the score and not to devise a better method of scoring buildings. We can make a note in our report that the scores have a suspect distribution, but our main focus is on predicting the score using the statistics.

Let us look at the distribution of the ‘*site EUI*’ variable.

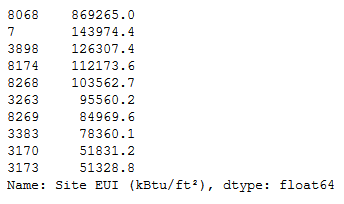




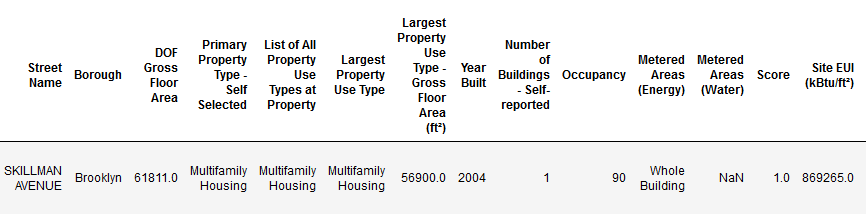
Well, this graph shows us another problem: outliers! The graph is incredibly skewed because of the presence of a few buildings with very high scores. It seems that we will have to take a slight detour to deal with the outliers. Let us look at the stats for this particular feature.





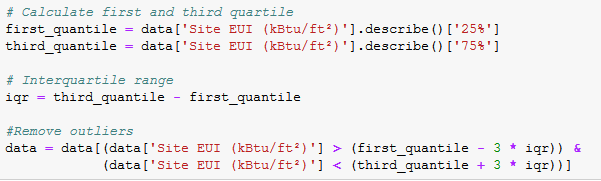


Wow! One building is clearly far above the rest

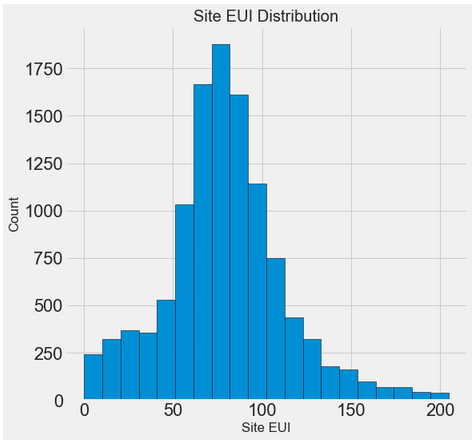


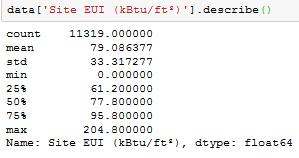
Looking at the data, it might be worthwhile for someone to follow up with this building owner! However, that is not our problem as statisticians, as we only need to figure out how to handle this information. Outliers can occur for several reasons: typos, malfunctions in measuring devices, incorrect units, or they can be legitimate but extreme values. Thus, the outliers can throw off a model, because they are not indicative of the actual distribution of data.

## Removing Outliers



Now, let’s check the distribution of the ‘*site EUI’* variable.





This plot looks a little less suspicious and is close to a normal distribution with a long tail on the right side (it has a positive skew).

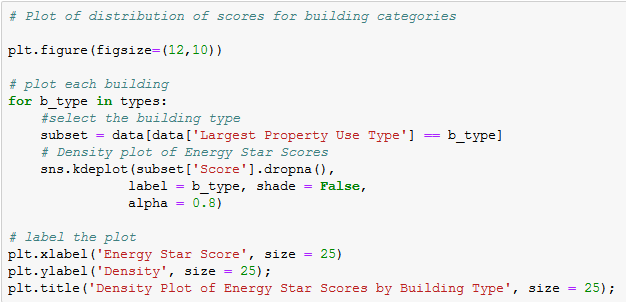
## Bivariate Analysis

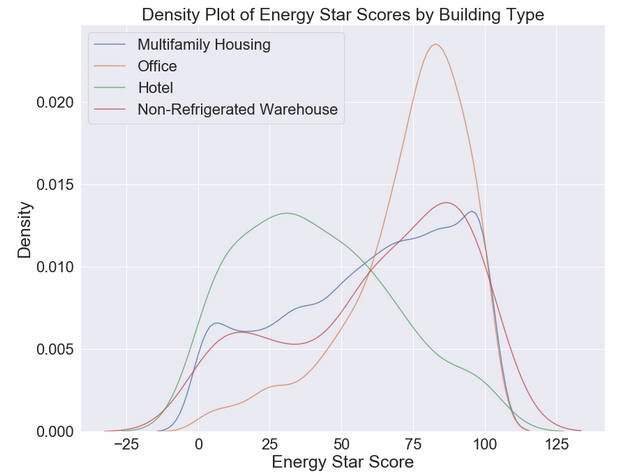
In order to look at the effect of categorical variables on the score, we can make a density plot, which is colored by the value of the categorical variable. Density plots also show the distribution of a single variable and can be thought of as a smoothed histogram. If we color the density curves by a categorical variable, this will show us the change in distribution based on the class.

The first plot we will make shows the distribution of scores by the property type. In order to not clutter the plot with large data, we will limit the graph to building types that have more than 100 observations in the dataset.



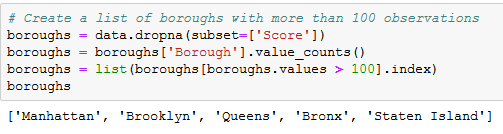


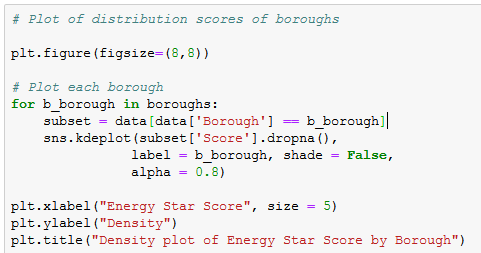


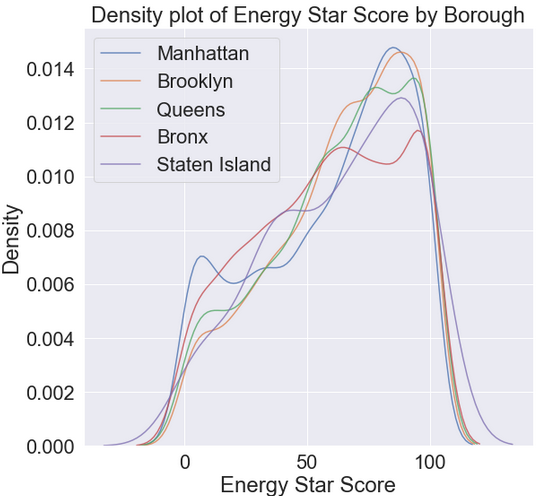


This graph informs us that we should include the property type for analysis, because this information can be useful for determining the score.

To examine another categorical variable, ‘*’borough’’*, we can make the same graph, but this particular one is time colored by the *borough*.

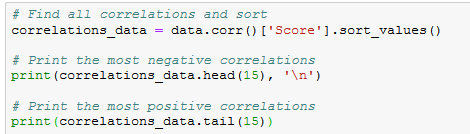




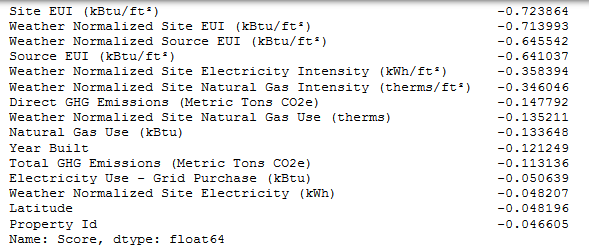


The borough of the building does not seem to make a significant difference in the distribution of the score in a line similar to the building type. Nonetheless, it might make sense to include the borough as a categorical variable for the final analysis.

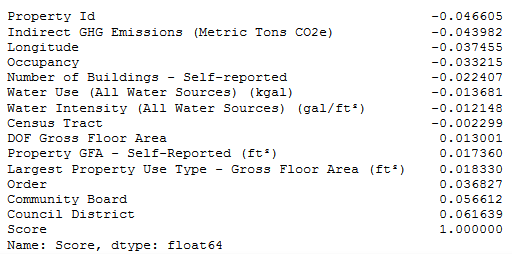
## Correlations between Features and Target



Top 15



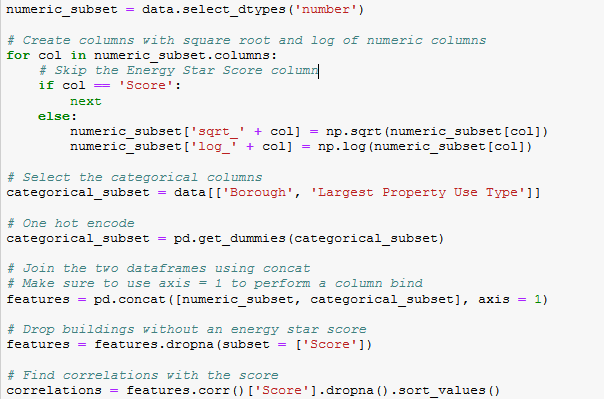
Bottom 15

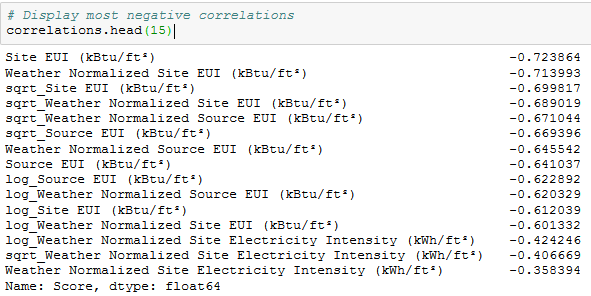


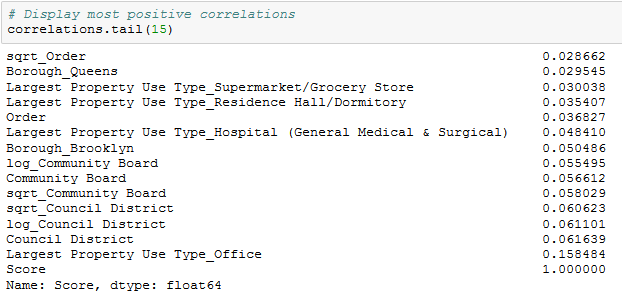
We can observe a number of strong negative correlations between the features and the target. The most prominent of these correlations with the score are the different categories of Energy Use Intensity (EUI), Site EUI (kBtu/ft²), and Weather Normalized Site EUI (kBtu/ft²). Generally, these categories slightly vary in terms of the process in which they are calculated. The EUI is the amount of energy utilized by a building divided by the square footage of the buildings (i.e., unit area) and is a measure of the efficiency of a building. The lower score indicates higher efficacy. Consequently, these correlations make sense: as the EUI increases, the Energy Star Score tends to decrease.

To account for the possible non-linear relationships, we can take square root and natural log transformations of the features and then calculate the correlation coefficients with the score. In this way, we try to capture any possible relationships between the *borough* and building type.

In the following code, we take the log and square root transformations of the numerical variables, one-hot encode the two selected categorical variables (building type and borough), calculate the correlations among all of the features and the score, and display the top 15 most positive and top 15 most negative correlations. This is a lot of work, but with pandas, it is a straightforward task through each step!



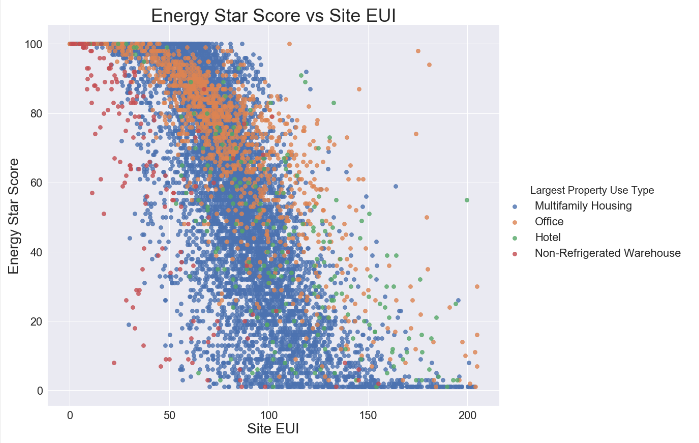




After transforming the features, we find the strongest relationships that relate to the Energy Use Intensity (EUI). We observe that the log and square root transformations did not have strong relationships. The positive linear relationships are not very strong, although we do see that a building type of office (Largest Property Use Type\_Office) is slightly positively correlated with the score. This variable is a one-hot encoded representation of the categorical variables for building type.

We can employ these correlations in order to perform the feature selection (coming up in future chapters). Right now, let us draw a graph of the most significant correlation (in terms of absolute value) in the dataset which is the *Site EUI* (kBtu/ft^2). We can color the graph by building type to show how it affects the relationship.

In order to visualize the relationship between two variables, we utilize a scatter plot. Moreover, we can also include the additional variables using aspects such as color or size of the markers. Here, we plot two numeric variables against one another and use a different color to represent a third categorical variable.

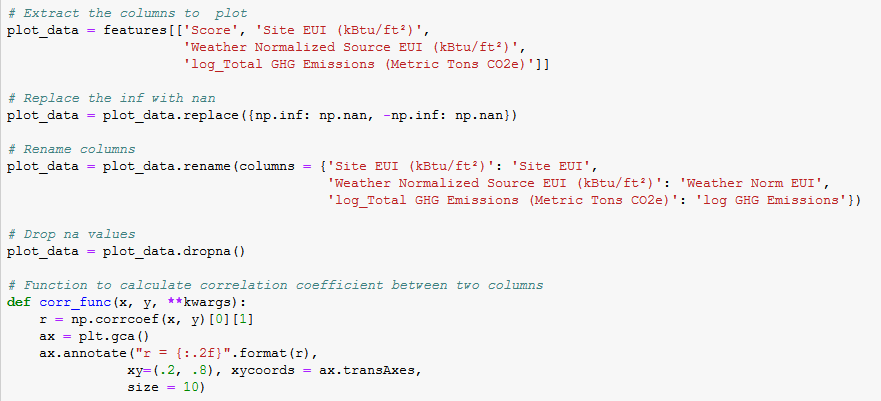


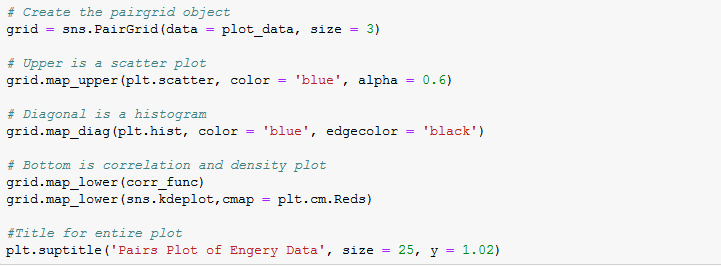
The graph shows a clear negative relationship between the Site EUI and the score. The relationship is not perfectly linear (it has a correlation coefficient of -0.7, but it does look like this feature will be important for predicting the score of a building.

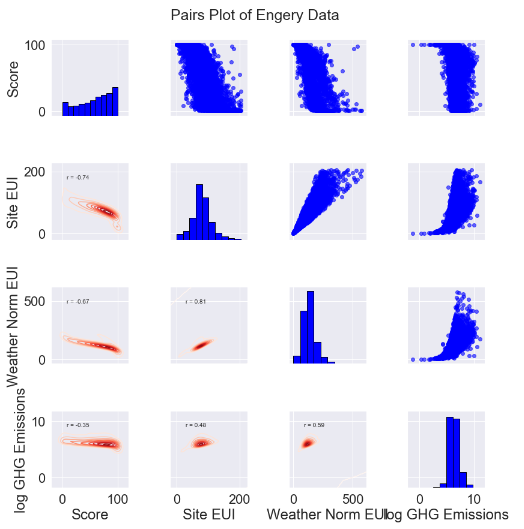
## Multivariate Analysis

As a final exercise for exploratory data analysis, we can make a pairs plot among several different variables. The pairs plot is a great way to examine multiple variables at once as it shows the scatterplots between pairs of variables and histograms of single variables on the diagonal axis.

Using the seaborn PairGrid function, we can map different plots for the three aspects of the grid. The upper triangle will have the scatterplots, the diagonal will show the histograms, and the lower triangle shows both the correlation coefficient between two variables and a 2-D kernel density estimate of the two variables.



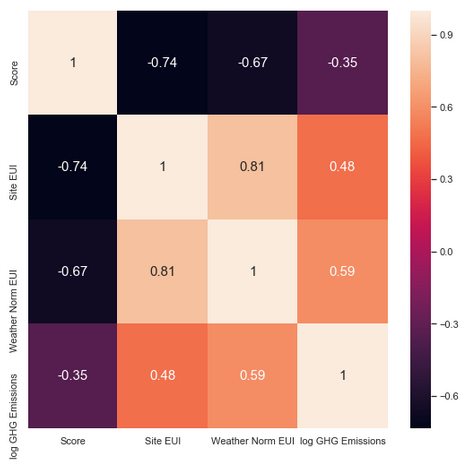




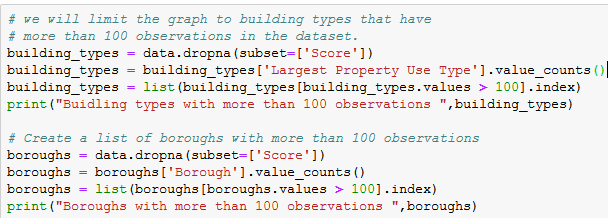
Here, we observe that all the three variables have a negative correlation with our target variable “Score”.

We also utilize the heat map to visualize the correlation between the continuous variables.



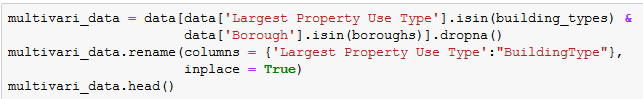


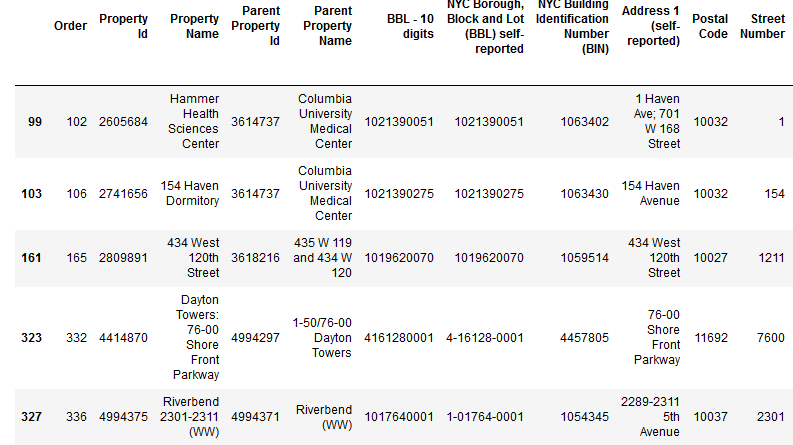
Let us understand the energy distribution score by grouping the data with categorical variables using facet-grid from the seaborn package.





We filter the dataset based on the building types and *boroughs*.









By looking at the graph, most of the multifamily buildings with high energy scores are located in Manhattan, Brooklynn, and Bronx *boroughs*.

The majority of the office and hotel properties are located in Manhattan. Furthermore, it can be observed that the office buildings have high-energy scores and hotel buildings have low-energy scores.

## Assessment

## Choose the appropriate option

1. **Which of the following analyses require more than two variables for analysis?**
   1. Univariate analysis
   2. Bivariate analysis
   3. Multivariate analysis
   4. None of the above
2. **Which of the following plots is used to study the relationship between two continuous variables?**
   1. Scatter plot
   2. Bar plot
   3. Histograms
   4. None of these
3. **Outliers are caused due to..?**
   1. Data entry
   2. Mistake in units
   3. Legitimate extreme value
   4. All of the above
4. **Univariate analysis is performed on..?**
   1. Single variable
   2. Two variable
   3. More than two variables
   4. None of the above
5. **What should be your next step if a variable has more than 80% missing values and the variable is not important for the analysis?**
   1. Drop the variable
   2. Impute the missing values
   3. Remove the observation with missing values
   4. None of the above

## Programming Assignment

Using the data in the below URL,

<https://www.kaggle.com/ruiromanini/mtcars>

1. Perform the univariate analysis on the mpg variable.
2. Check the correction between mpg and rest of the continuous variables.
3. Perform multivariate analysis by grouping the data with am, gear, and cyl variables.

Solutions: Refer to page 210

## Solutions for Assessment

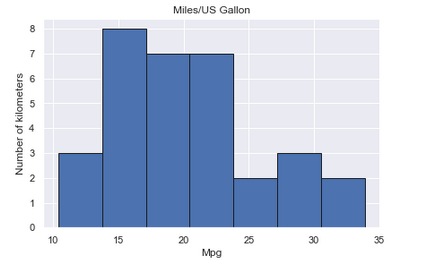
## Choose the appropriate options

1. C
2. A
3. D
4. A
5. A

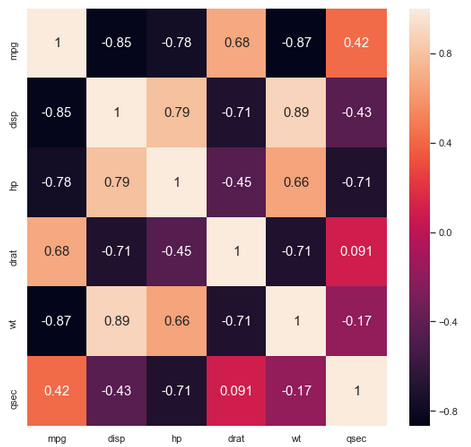
## 

## Programming Assignment Solution

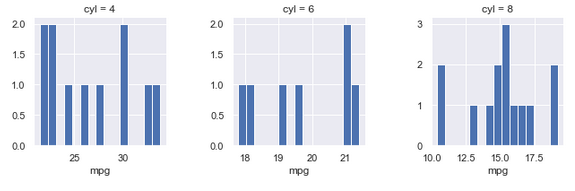
Task 1)

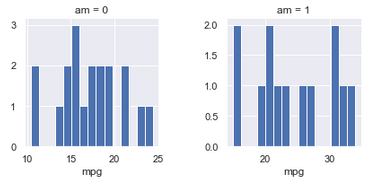


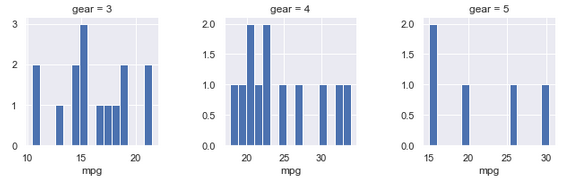
Task 2)



Task 3)







# CHAPTER 6: Regression Analysis - Part I

## Theory

In the previous chapter, we worked on data cleaning and wrangling. In this chapter, we will learn the core topic of model building for making predictions.



Regression analysis is an important tool for analyzing and modeling data. Here, we fit a line through the data points, in such a manner that the difference between the distances of the actual data points from the plotted line is minimal.

## The use of Regression Analysis

Regression analysis investigates the relationship between two or more variables. As an example, consider the following:

Let’s suppose that we develop an application which predicts the chances of admission of a student to a foreign university.

The benefits of using Regression analysis for calculating these chances are as follows:

* It provides the significant relationships between the label (dependent variable) and the feature (independent variable).
* It shows the extent of the impact of multiple independent variables on the dependent variable.
* It can also quantify these effects even if the variables are on a different scale.

These features enable a data analyst to find the best set of independent variables for making the predictions.

## Linear Regression

Linear Regression is one of the most fundamental and widely-known Machine Learning Algorithms, and is often the first regression analysis performed on the related dataset.

Building blocks of a Linear Regression Model are:

* Discrete/Continuous independent variable.
* A best-fit regression line.
* Continuous dependent variable*. i.e*., A Linear Regression model predicts the dependent variable using a regression line based on the independent variables. Linear Regression can be modeled as follows:

**Y = MX + C**



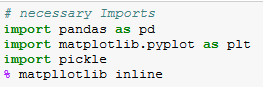
Where,

* Y = Dependent variable
* M = Slope
* X = Independent variable
* C = Constant or Intercept

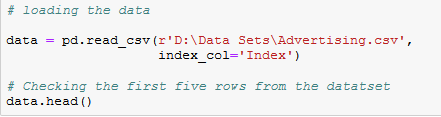
## Problem Statement

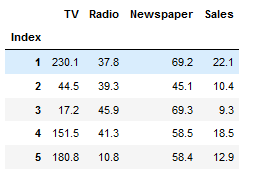
This data is about the amount spent on advertising certain products through various media channels like TV, radio, and newspaper. The goal is to predict how the expense on each channel affects the sales and is there a way to optimize the sales?

Importing the necessary packages.



Loading and exploring the data.





What are the **features**?

1. TV: Dollars spent on TV ads for a single product in a given market (expressed in multiple of thousands).
2. Radio: Dollars spent on Radio ads (expressed in multiple of thousands).
3. Newspaper: Dollars spent on Newspaper ads (expressed in multiple of thousands).

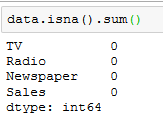
What is the **response**?

* **Sales**: sales of a single product in a given market (in thousands of widgets).

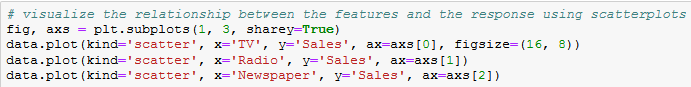
Dimensions of the data

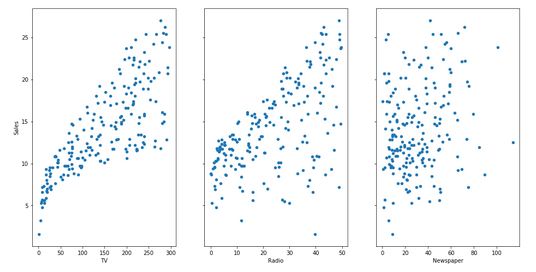


Find the missing values from different columns:

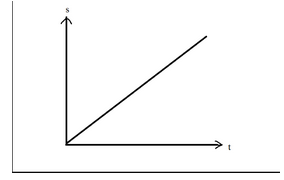


Let us showcase the relationship between the feature and target variables:





From the relationship diagrams above, it can be observed that there is a linear relationship of the features such as TV ad, radio ad with the sales. A linear relationship typically looks like:



Hence, we can build a model using the Linear Regression Algorithm.



The mathematical equation is given as:

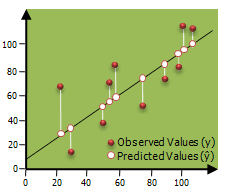
***𝑦* =*𝛽*0 + *𝛽*1*𝑥***

What do the terms in the equation represent?

* **Y:** Response or Target Variable
* **X:** Feature Variable
* **𝛽1:** Coefficient of X
* **𝛽0:** Intercept

Where 𝛽0 and 𝛽1 are the **model coefficients**. To create a model, we must "*learn*" the values of these coefficients. Moreover, once we have these values, we can use the model to predict the Sales!

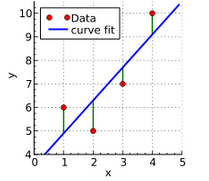
## Ordinary Least Squared Estimation



* In a given two-dimensional space, an infinite number of lines can be plotted through the scatter diagram between the two variables.
* Calculate the distance between the observed values and predicted values, which are called Errors or Residuals.
* Errors can be positive and negative, therefore square the error term (squared Error) to make them all positive.
* Repeat the process with an infinite number of lines and identify the line with minimum sum of squared error.
* Hence, the name “*Ordinary least Squared*”.

## The Mathematics involved

Take a quick look at the plot created. Next, consider the X and Y coordinates of each point. Now draw an imaginary line between the points and the current "best-fit" line. We will call the distance between each point and the current best-fit line as D. To get a quick image of what we're trying to visualize, take a look at the picture below:



What elements are present in the diagram?

* The red points are the observed values of X and Y.
* The blue line is the least square line.
* The green lines are the residual values, which are the distances between the observed values and the least squared line.

Now label each green line as having a distance D, and each red point as having a coordinate of (X, Y). Next, we can define our best fit line as the lines having the property were:

𝐷21+𝐷22+𝐷23+𝐷24+....+𝐷2𝑁

So how do we find this line? The least-square line approximating the set of points:

**(𝑋,𝑌)1,(𝑋,𝑌)2,(𝑋,𝑌)3,(𝑋,𝑌)4,(𝑋,𝑌)5,**

Has the equation:

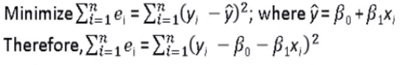
***𝑦* =*𝛽*0 + *𝛽*1*𝑥***

Basically, it is just a rewritten form of the standard equation for a line:

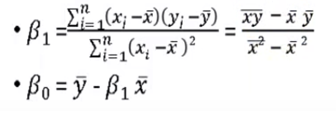
**Y = MX + C**

## Derivation of OLS by Minimizing Errors

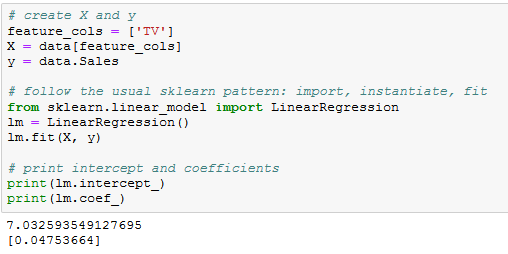
Minimize the sum of squared error term by substituting as below:



Solving the above equation by calculus – partial differencing by ***𝛽*0 and *𝛽*1** respectively and solving for the two variables, we get the following equations,



Building Simple Linear Regression Model to predict the sales based on TV ads,



## Interpreting the model

How do we interpret the coefficient for spending on TV ads (β1)?

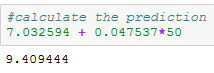
* A “unit” increase in spending on a TV ad is **associated with** a 0.04753 “unit” increase in sales.
* Or, an additional $1,000 on TV ads is translated to an increase in sales by $47.53.

## Prediction using the model

If the expense on a TV ad is $50000, what will be the sales prediction for that market?

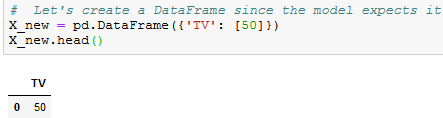
***𝑦* =*𝛽*0 + *𝛽*1*𝑥***

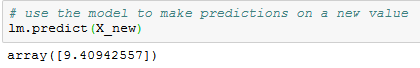
Y = 7.032594 + 0.047537 \* (50)



Thus, we can predict Sales of 9,409 widgets in that market.

Let’s do the same calculation using the code.

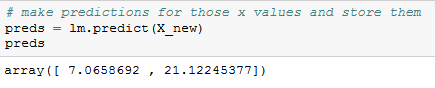


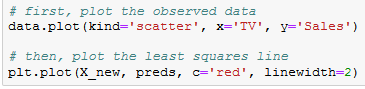


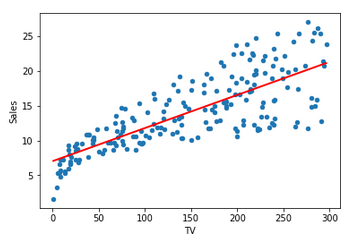
Plotting the least Squares Line:











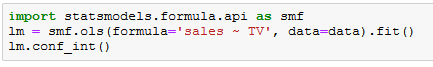
## Model Confidence

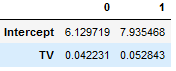
**Question:** Is the linear regression a low bias/high variance model or a high bias/low variance model?

**Answer:** It is a High bias/low variance model. Even after repeated sampling, the best fit line will stay roughly in the same position (low variance), but the average of the models created after repeated sampling does not do a great job in capturing the perfect relationship between the two variables (high bias). Low variance is helpful when we don't have less training data!

If the model has calculated a 95% confidence interval for our model coefficients, it can be interpreted as follows: If the population, from which this sample is drawn, is **sampled 100 times**, then approximately **95 (out of 100) of those confidence intervals** shall contain the "*true*" coefficients.

In the coming sections, we discuss more about bias and variance in detail.





## Hypothesis Testing and p-values

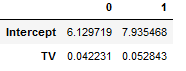
**Hypothesis testing** is closely related to the confidence intervals. We start with a **null hypothesis** and an **alternate hypothesis** (*i.e.,* opposite to the *null*). Subsequently, we check whether the data **rejects the null hypothesis** or **fails to reject the null hypothesis**.

The conventional hypothesis test is as follows:

* **Null hypothesis:** No significant relationship exists between the TV advertisements and the Sales (and hence, *𝛽*1 equals zero).
* **Alternative hypothesis:** A significant relationship exists between the TV advertisements and Sales (and hence, *𝛽*1 is not equal to zero).

How do we test this potential relationship? We reject the null hypothesis (and thus believe the alternative hypothesis) if the 95% confidence interval **does not include zero**. The **p-value** represents the probability that the coefficient is actually zero.









If the 95% confidence interval **includes zero**, then the p-value for that coefficient will be **greater than 0.05**. If the 95% confidence interval **does not include zero**, the p-value for the coefficient will be **less than 0.05**.

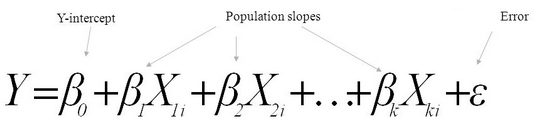
Thus, a p-value of less than 0.05 shows that the relationship between the two features in consideration is statistically significant. Conventionally, a cut-off p-value of 0.05 is used for such analysis.

In this case, the p-value for TV ads is far smaller than 0.05, which means that a statistically significant relationship exists between the TV advertisements and Sales.

Note that we generally ignore the p-value for the intercept.

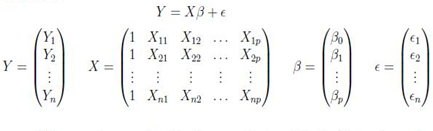


Till now, we have studied the models based on only one feature. Now, we’ll include models with multiple features and investigate the potential relationship between those features and the target column. This is called **Multiple Linear Regression**.

****

**Estimation of model parameters**

Consider the model where,



The columns of X are each covariate for the n number of patients, with the first column having all 1’s to include the intercept in the model.

Based on this model, we get the following expansion for the first subject:



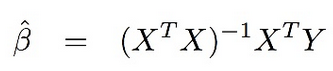
Then using the matrix calculus, we can find that the least square estimate for β is given by,



Hence, the least squares regression line is:



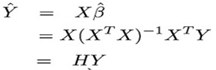
The beta values are obtained by calculating the below equation:



Note:

1. (XTX)-1 should be a non-singular matrix, otherwise, we cannot calculate the value of beta.
2. Beta cannot be calculated if columns are linear combinations of each other.

We know that,



H is called a Hat matrix

Properties of Hat matrix:

1. H is symmetric; HT = H
2. H is idempotent; HH = H

## Assumptions of OLS Regression

Regression is a parametric approach. ‘*Parametric’* means it makes assumptions about data for the purpose of analysis. Due to its parametric side, regression is restrictive in nature. It fails to deliver trustworthy results from a dataset, which does not fulfill its assumptions. Therefore, for a successful regression analysis, it is essential to validate these assumptions.

Thus, how would you check (validate) if a data set follows all the assumptions for regression? You check it using the regression plots (explained below) along with some statistical tests.

Let’s look at the important assumptions in regression analysis:

1. There should be a linear and additive relationship between the dependent (response) variable and independent (predictor) variable(s). A linear relationship suggests that a change in response Y due to a constant change in X1 is the same, regardless of the value of X1. An additive relationship suggests that the effect of X1 on Y is independent of other variables.
2. There should be no correlation between the residual (error) terms: An absence of this phenomenon is known as *Autocorrelation*.
3. The independent variables should not be correlated: An absence of this phenomenon is known as *multicollinearity*.
4. The error terms must have a constant variance. This phenomenon is known as *homoscedasticity*. The presence of non-constant variance is referred to as *heteroskedasticity*.
5. The error terms must be normally distributed.

## What if these assumptions are violated?

Let’s dive into specific assumptions and learn about their outcomes if the assumptions are violated:

1. **Linear and Additive:**  If you fit a linear model to a non-linear, non-additive dataset, the regression algorithm will fail to capture the trend mathematically, producing an inefficient model. This will also result in erroneous predictions on an unseen dataset.

**How to check for this factor:** Look for residual vs fitted value plots (explained below).

1. **Autocorrelation:** The presence of correlation in error terms drastically reduces the accuracy of a model. This usually occurs in time series models, where the next instant is dependent on the previous instant. Moreover, if the error terms are correlated, the estimated standard errors tend to underestimate the true standard error.

If this happens, it narrows down the confidence intervals and prediction intervals. Narrower confidence interval means that a 95% confidence interval would have lower probability than 0.95 of containing an actual value of coefficients. Let’s understand the narrow prediction intervals with an example:

For example, the least square coefficient of X1 is 15.02 and its standard error is 2.08 (without autocorrelation). But in the presence of autocorrelation, the standard error is reduced to 1.20. As a result of this reduction, the prediction interval narrows down to (13.82, 16.22) from (12.94, 17.10).

Additionally, the lower standard errors causes the associated p-values to be lower than actual. This can result in the incorrect conclusion that a certain is statistically significant.

**How to check for this factor:** Look for Durbin – Watson (DW) statistics. It must lie between 0 and 4. A DW = 2 implies no autocorrelation. A DW = 0 < DW < 2 implies a positive autocorrelation while DW = 2 < DW < 4 indicates a negative autocorrelation. Additionally, you can also examine the residual vs time plots and look for the seasonal or correlated patterns in residual values.

1. **Multicollinearity:** This phenomenon exists when the independent variables are found to be moderately or highly correlated. In a model with correlated variables, it is hard to figure out the true relationship of predictors with the response variables. In other words, it is difficult to find out which variable is actually contributing to predict the response variable.

Additionally, with the presence of correlated predictors, the standard errors tend to increase. And, with large standard errors, the confidence interval becomes wider which leads to less precise estimates of slope parameters.

Further, when predictors are correlated, the estimated regression coefficient of a correlated variable depends on the availability of other predictors in the model. If this happens, you’ll make an incorrect conclusion that a given variable strongly or weakly affects the target variable. In such cases, dropping a single correlated variable from the model can change the estimated regression coefficients, which will induce errors.

**How to check for this factor:** You can use scatter plot to visualize the correlation effect among variables. Further, you can also use the VIF factor. VIF value ≤ 4 suggests no multicollinearity whereas a value of ≥ 10 implies strong multicollinearity. Above all, a correlation table should also solve the purpose.

1. **Heteroscedasticity:** The presence of non-constant variance in the error terms results in heteroscedasticity. Generally, non-constant variance arises in the presence of outliers or extreme leverage values. This means that these values get higher than the expected weightage, which disproportionately influences the model’s performance. When this phenomenon occurs, the confidence interval for prediction involving other samples tends to be unrealistically wide or narrow.

**How to check for this factor**: You can look at the plot for the residual vs fitted values. If a heteroscedasticity exists, the plot would exhibit a funnel shape pattern. You can also use *Breusch-Pagan* / *Cook–Weisberg* test or White general test to detect this phenomenon.

1. **Non-normal Distribution of error terms:** If the error terms are non-normally distributed, confidence intervals becomes too wide or narrow. Once the confidence interval becomes unstable, it leads to difficulty in estimating the coefficients based on minimization of least square values. Presence of non – normal distribution suggests that there are a few unusual data points, which must be closely studied to make a better model.

**How to check for this factor:** You can look at the QQ plot (shown below). You can also perform statistical tests of normality such as the Kolmogorov-Smirnov test, Shapiro-Wilk test.

## Interpretation of Regression Plots

So far, we’ve learned about the important assumptions for regression and the methods to undertake, if those assumptions get violated.

But that is not the end. Now, you should know the solutions to tackle the violation of these assumptions. In this section, I have explained the 4 regression plots along with the methods to overcome the limitations on assumptions.

## Residual vs Fitted Values

## 

This scatter plot shows the distribution of residuals values (errors) vs fitted values (predicted values). It is an important plot to learn for everyone. It reveals several useful insights including outliers into regression models. The outliers in this plot are labeled by their observation numbers which make them easy to detect.

There are two major facts to learn from this:

* If a potential pattern (maybe, a parabolic shape) exists in this plot, consider it as a sign of non-linearity in the data. It means that the model doesn’t capture linear effects.
* If a funnel shape is evident in the plot, consider it as the signs of non-constant variance *i.e.,* heteroscedasticity.

**Solution:** To overcome the issue of non-linearity, you can do a nonlinear transformation of predictors such as log (X), √X or X² and transform the dependent variable. To overcome heteroscedasticity, a possible way is to transform the response variable such as log(Y) or √Y. You can also use a weighted least square method to tackle heteroscedasticity.

## Normal Q-Q Plot

## 

This q-q or quantile-quantile is a scatter plot which helps us validate the assumption of normal distribution in a data set. This plot can be used to infer if the data comes from a normal distribution. In such case, the plot would show a fairly straight line. Absence of normality in the errors can be seen with deviation in the straight line.

If you are wondering what is ‘*quantile’*, here is a simple definition: Quantiles as points in your dataset below which a certain proportion of data falls. Quantile is often referred to as percentiles. For example: when we say the value of the 50th percentile is 120, it means half of the data set lies below 120.

**Solution:** If the errors are not normally distributed, non-linear transformation of the variables (response or predictors) can bring improvement in the model.

## Scale Location Plot

## 

This plot is also utilized to detect *homoscedasticity* (assumption of equal variance). It shows how the residuals are spread along a range of predictors. It is similar to residuals vs fitted value plot, except that it uses standardized residual values. Ideally, there should be no discernible pattern in the plot. This implies that the error values are normally distributed. However, if the plot shows a discernible pattern (probably a funnel shape), it would imply a non-normal distribution of errors.

**Solution:** Follow the solution for heteroskedasticity, as given in plot 1.

## Residuals vs Leverage Plot

## 

It is also known as Cook’s Distance plot, and attempts to identify the points which have a higher influence on the dataset than the other points. Such influential points tend to have a sizable impact on the regression line. In other words, adding or removing such points from the model can completely change the model statistics.

But, can these influential observations be treated as outliers? This question can only be answered after carefully evaluating the data. Therefore, in this plot, the large values marked by cook’s distance might require further investigation.

**Solution:** For influential observations which are outliers, you can omit them if they are few. Alternatively, you can scale down the outlier observation with maximum value in the data or else treat those values as missing values.

## Model Evaluation Metrics

Model evaluation is a very important procedure in data analysis. It helps you to understand the performance of your model and simplify its presentation. Several evaluation metrics are available but only some of them are suitable for use in regression analysis.

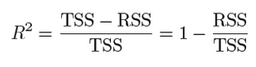
There are 5 major metrics for model evaluation in regression analysis:

1. R Square
2. Adjusted R Square
3. Mean Square Error (MSE)
4. Root Mean Squared Error (RMSE)
5. Mean Absolute Error

## R Square/Adjusted R Square

R Square measures the amount of variability independent variables that can be explained by the model. It is the square of Correlation Coefficient (R), hence called R Square. It takes the form of a proportion --- the proportion of variance explained --- therefore, its value is always between 0 and 1.

For example, **R2** statistic = 0.75, it says that our model fits 75% of the whole dataset. Similarly, if it is 0, it means none of the data points is being explained and a value of 1 represents 100% data explanation. Mathematically, **R2** statistic is calculated as:



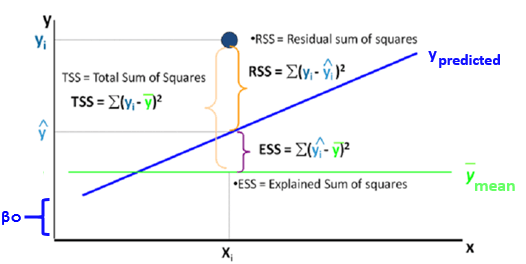
Where **RSS** is the Residual Sum of Squares and is given as:



RSS is the residual (error) term, whereas, TSS is the Total Sum of Squares and given as:



TSS is calculated when we consider the line passing through the mean value y, to be the best fit line. This calculation is very similar to the calculation for RSS.



The closer the value of R2 is to 1, the better the model fits our data. If the value of R2 is low (which is a possibility), then the model is so bad that it is performing even worse than the average best fit.

## Adjusted R Square

R Square is a good measure to determine how well the model fits the dependent variables. However, it does not take into consideration the overfitting problem. If your regression model has many independent variables, because it is too complicated, it may fit very well to the training data but perform badly for testing data. Adjusted R Square is introduced to overcome this problem because it penalizes additional independent variables added to the model and adjust the metric to prevent the overfitting issues.

Mathematically, it is calculated as:



Where:

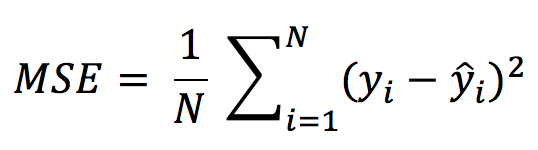
* **R2**: Sample R Square
* **N**: Total Sample Size
* **P**: Number of Independent variables

## Mean Square Error (MSE)

While R Square is a relative measure of how well the model fits the dependent variables, Mean Square Error (MSE) is an absolute measure of the goodness of fit.

MSE is calculated by the sum of prediction error, which is the real output minus the predicted output, and then divided by the number of data points. It gives an absolute number on how much your predicted results deviate from the actual number. You cannot interpret much insight from one single result, but it gives you a real number to compare against other model results, and helps you to select the best regression model.

Mathematically, it is calculated as:



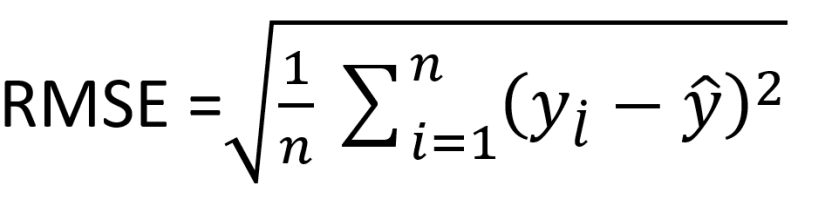
Where:

* **Yi**: actual value
* **Yi hat**: predicted value
* **N**: No. of observations

## Root Mean Square Error (RMSE)

Root Mean Square Error (RMSE) is the square root of MSE. It is used more commonly as compared to MSE, because sometimes MSE values can be too big for easy comparisons. Secondly, MSE is calculated by the square of error, and thus it brings to the same level of prediction error and makes it easier for interpretation.

Mathematically it is calculated as:



Where:

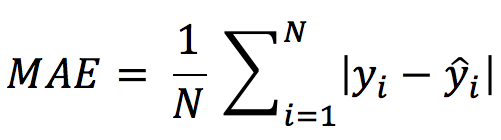
* **Yi**: actual value
* **Yi hat**: predicted value
* **N**: No. of observations

## Mean Absolute Error

Mean Absolute Error (MAE) is similar to Mean Square Error (MSE). However, instead of the sum of square error in MSE, MAE is taking the Sum of Absolute value of error.

Compared to MSEand RMSE, MAE is a more direct representation of the sum of error terms. MSE gives larger penalization to big prediction errors by square them while MAE treats all errors the same way.

Mathematically, MAE is calculated as:



Where,

* **Yi**: actual value
* **Yi hat**: predicted value
* **N**: No. of observations

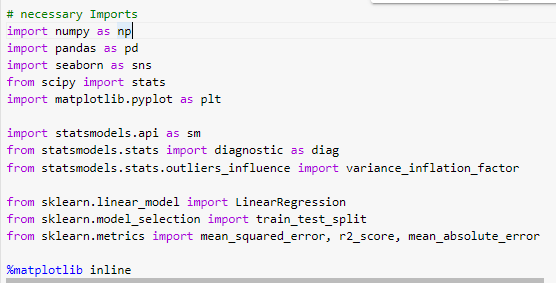
## Overall Recommendation/Conclusion

R Square/Adjusted R Square are better applied to explain the model to other people because you can explain the number as a percentage of the output variability. MSE, RMSE or MAE are more useful for comparisons among different regression models.

## Practical Implementation

## Import our libraries

The first thing we need to do is to import the libraries we will be using in this case study.

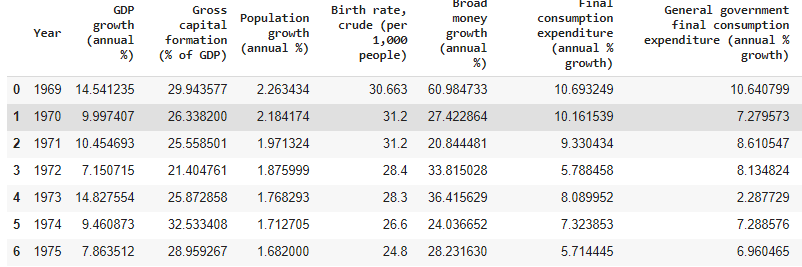


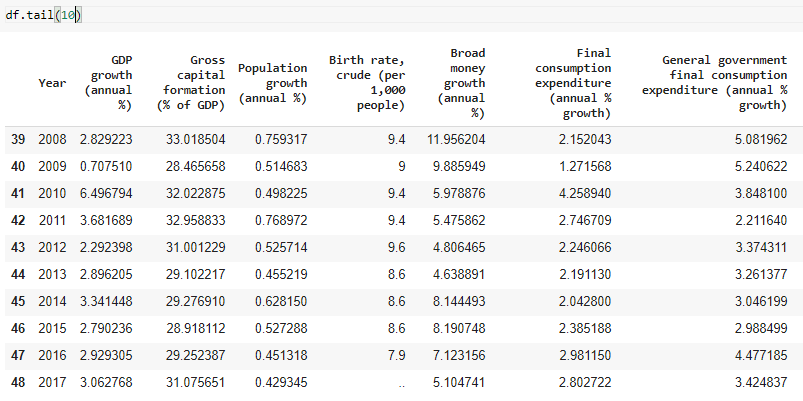
## Load the Data into Pandas

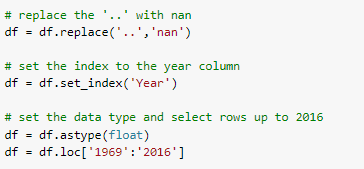
This dataset was downloaded from the World Bank website; if you intend to visit the website yourself, you can visit the following link. There is a tremendous amount of data available for free that can be used across a wide range of models.

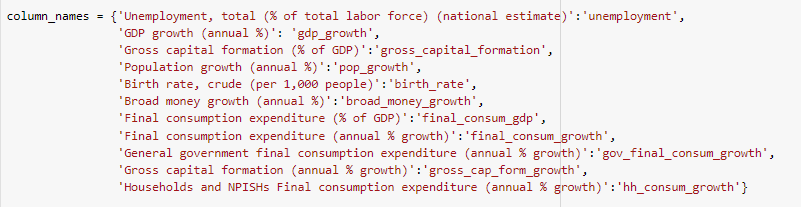
Link: <https://data.worldbank.org/>

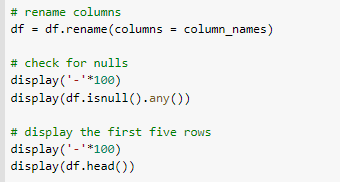




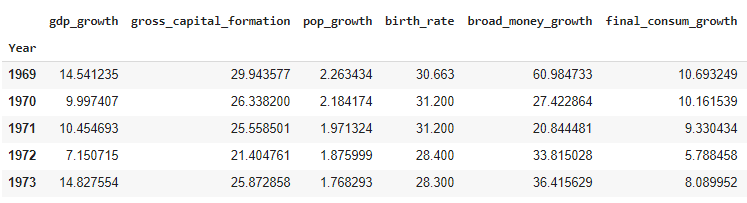












We can observe that a few missing values are represented as (..) in the dataset, we need to replace all the (..) values with “*nan*” as these represent the missing values in our dataset.

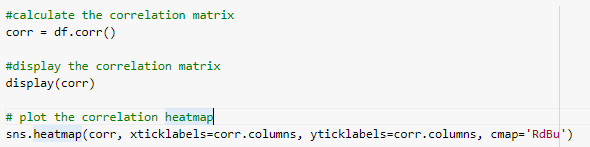
We set the index of our data frame using the set\_index() function to the Year column. This step will make the data selection easy. After we have defined the index, we convert the entire data frame to a float data type and then select the years ranging from 1969 to 2016. These years are selected because they do contain missing values.

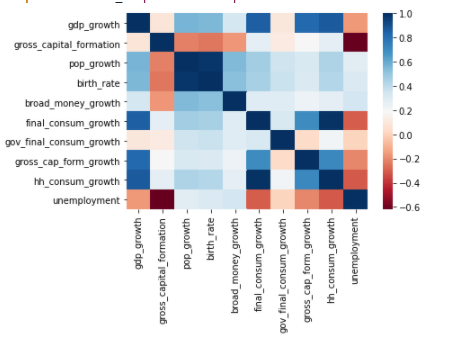
To make the selection of the columns a little easier, we have renamed all the columns by creating a dictionary where the keys represent the old column names and the values associated with those keys are the new column names.

Finally, I have checked for potential missing values using isnull().any(), which will return *true* for a given column if any values are missing, and then printed the head of the data frame.

## Check for Multicollinearity

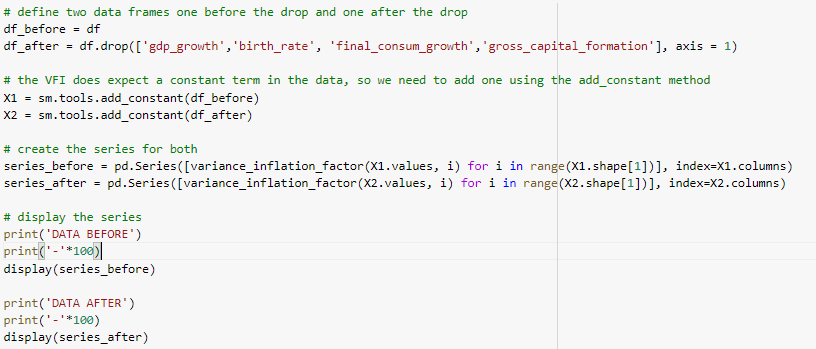
The first thing to do after loading our data is to validate the assumptions of our model; in this case, we will be checking for multicollinearity.

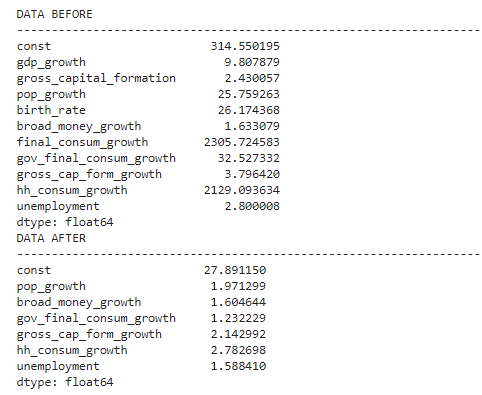




Looking at the heatmap along with the correlation matrix, we can identify a few highly correlated variables. For example, if you look at the correlation between birth\_rate and pop\_growth, it shows a maximal value of 0.98. This is an extremely high correlation and requires removal. Logically, it makes sense that these two are highly correlated; if you are having more babies, then the population should be increasing.

However, we should be more systematic in our approach of removing highly correlated variables. One method for such a purpose is the Variance\_Inflation\_Factor, which is a measure of how much a particular variable is contributing to the standard error in the regression model. When significant multicollinearity exists, the variance inflation factor will be huge for the variables in the calculation.



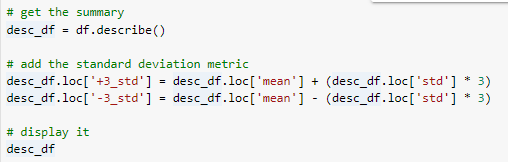


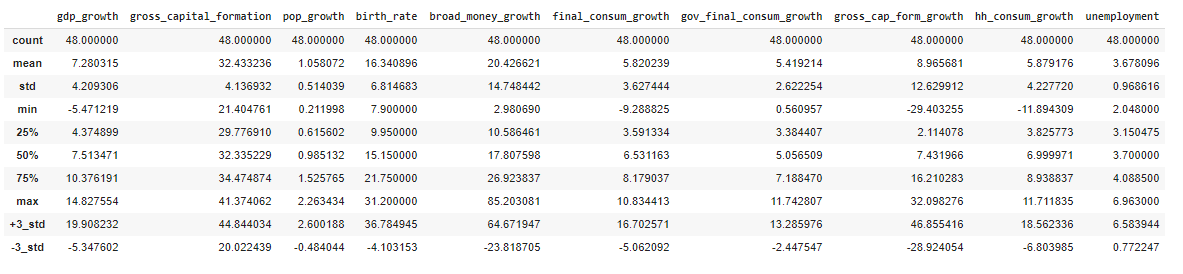
Looking at the above data, we get some confirmation about our suspicion. It makes sense to remove either birth\_rate or pop\_growth, and some of the consumption growth metrics. Once we remove those metrics and recalculate the VIF, we get a passing grade and can move forward.

## Describe the Dataset

Before we get to an in-depth exploration of the data, or even building the model, we should investigate the data a little more and see how the data is distributed and look for potential outliers. I will be adding a few more metrics to the summary data frame, so that it includes a metric for three standard deviations below and above the mean.

I will store my information in a new variable called *desc\_df*.





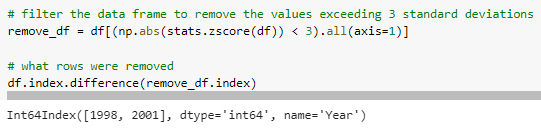
It is worth mentioning that we only have 50 observations, but 6 (minus the 3 we dropped) exploratory variables. Many people would argue that we need more data to have these many exploratory variables and this indeed is a correct statement. Generally, we should aim for at least 20 instances for each variable; however, the opinions vary and some statisticians argue that only 10 instances would also suffice.. Regardless, we only end up with 4 exploratory variables, so that we will satisfy that rule.

Looking at the data frame up above, a few values are standing out, for example, the maximum value in the *broad\_money\_growth* column is almost four standard deviations above the mean, which is such an enormous value qualifies as an outlier.

## Filtering the Dataset

To drop or not to drop a value, that is the question. Generally, if we believe that there is an erroneous entry in the data, we should remove it. However, in this situation, the values that are being identified as outliers are correct values and are not errors. Both of these values were produced during specific moments in time. The one in 1998 was right after the Asian Financial Crisis, and the one in 2001 was right after the DotCom Bubble, so it is entirely conceivable that these values were produced in extreme albeit rare conditions. **For this reason, I will NOT be removing these values from the dataset as they recognize actual values that took place.**

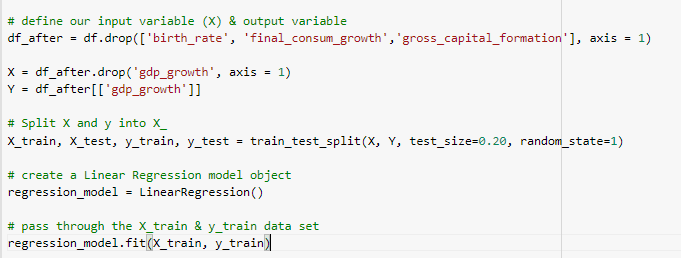
Imagine if we wanted to remove the values that have an amount exceeding three standard deviations. How will we approach this? Well, if we leverage the numpy and the scipy modules, we can filter out the rows by using the *stats.zscore* function. The Z-score is the distance of a datapoint in terms of the number of standard deviations from the mean. Hence, if it is less than 3 we keep it in the dataset, otherwise we drop it. From here, I have also provided a way to identify what rows were removed by using the *index.difference function*, which will show the difference between the two datasets.



## Build the model

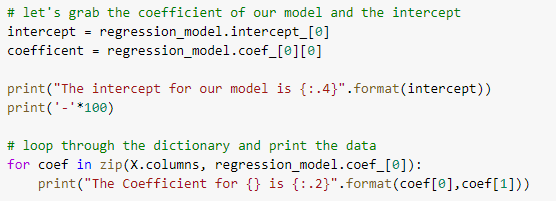
Now that we have loaded, cleaned, and explored the data, we can proceed to the next part, which is building the model. The first thing we need to do is, to define our exploratory and our explanatory variables. From here, let’s split the data into training and testing sets; a healthy ratio is 20% testing and 80% training but a 30% 70% split also works.

After splitting the data, we will create an instance of the linear regression model and pass through the X\_train and y\_train variables using the *fit( )* function.



## Exploring the output

With the data now fitted to the model, we can explore the output. The first thing we should do is to look at the intercept of the model, and then print out each coefficient value of the model. I print everything out using a loop to make it more efficient.

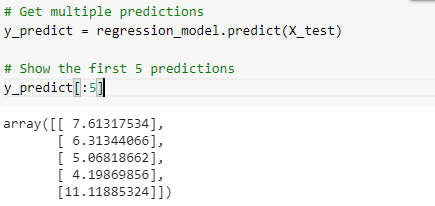




The intercept term is the value of the dependent variable, when all the independent variables are equal to zero. For each slope coefficient, it is the estimated change in the dependent variable for a one unit change in that particular independent variable, holding the other independent variables constant.

For example, if all the independent variables are equal to zero, the gdp\_growth would be 2.08%. If we look at the gross\_cap\_form\_growth, while keeping other independent variables constant, we would say that a 1 unit increase in gross\_cap\_form\_growth would lead to a 0.14% increase in GDP growth.

We can also make predictions using our newly trained model. The process is simple; we call the predict method and then pass through some values. In this case, we have some values predefined with the x\_test variable, so we will pass that through. Once we do that, we can select the predictions by slicing the array.



## Evaluating the model

## Using the Statsmodel

For diagnosing the model easier, we will, from now on be using the statsmodel module. This module has built-in functions that will help in fast and easy calculations of metrics. However, we will need to "*rebuild*" our model using the statsmodel module. We do this by creating a constant variable, call the *OLS()* method and then the *fit()* method. Now we have a new model, and the first thing we need to do is to ensure that the assumptions of our model hold. This means that we should check the following:

* Regression residuals must be normally distributed.
* The residuals are homoscedastic.
* Absence of multicollinearity (we did this above).
* No Autocorrelation.

## Checking for HeteroScedasticity

To check for heteroscedasticity, we can leverage the *statsmodels.stats.diagnostic* module. This module gives us a few test functions we can run, the Breusch-Pagan and the White test for heteroscedasticity. The **Breusch-Pagan is a general test for heteroscedasticity, while the White test is a unique case.**

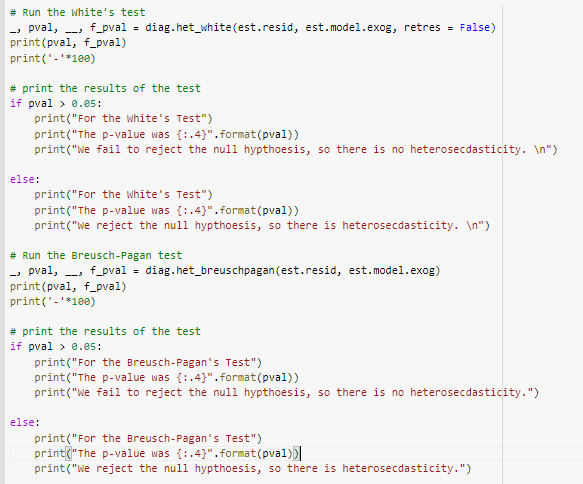
* The null hypothesis for both the White’s test and the Breusch-Pagan test states that the variances for the errors are equal:

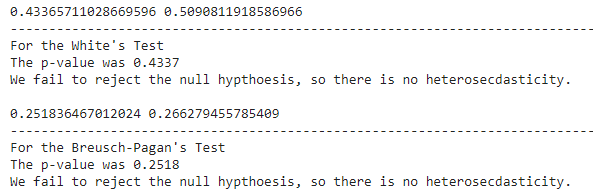
**H0 = σ2i = σ2**

* The alternate hypothesis (the one you’re testing), states that the variances are not equal:

**H1 = σ2i ≠ σ2**

We aim to fail to reject the null hypothesis, with a target high p-value because that implies that we found no heteroscedasticity in our dataset.





## Checking for Autocorrelation

We will go to our favorite *statsmodels.stats.diagnostic* module, and utilize the *Ljung-Box* test for no autocorrelation of residuals. Here:

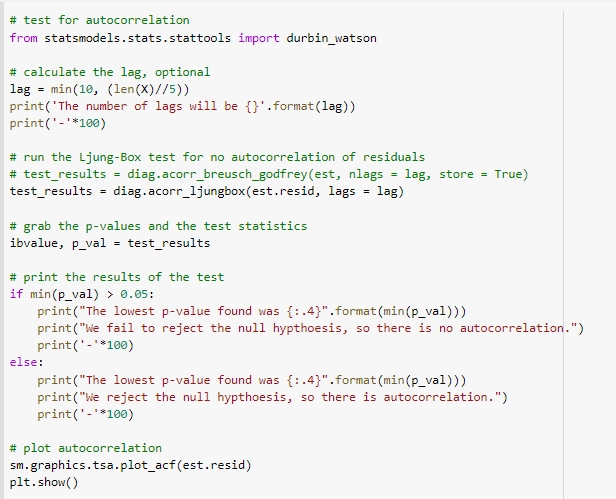
**H0: The data is random.**

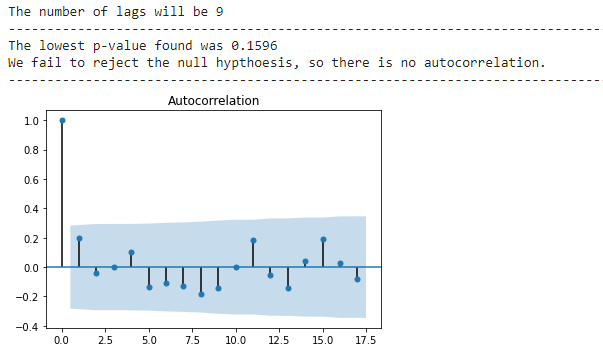
**Ha: The data are not random.**

That means that we want to fail to reject the null hypothesis, or specifically we want a large p-value because it would imply having no autocorrelation. To use the *Ljung-Box* test, we will call the *acorr\_ljungbox* function, pass through the *est.resid*and then define the lags.

The lags can either be calculated by the function itself, or we can calculate them manually. If the function handles it, the max lag will be min((num\_obs // 2 - 2), 40), however, there is a rule of thumb that for the non-seasonal time series, the lag is min(10, (num\_obs // 5)).

Moreover, we can also visually check for the autocorrelation by using the *statsmodels.graphics* module to plot a graph of the autocorrelation factor.





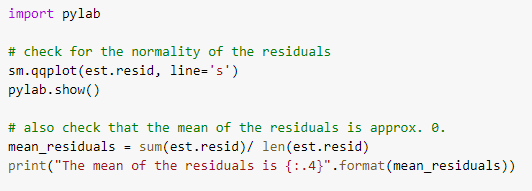
## Checking for Normally Distributed Residuals

This is an easy check and is visually possible. It **requires using a QQ pplot, which helps us to assess if a dataset plausibly came from some theoretical distribution such as a Normal or exponential.** Remember that it is just a visual check and not an air-tight proof, so it is somewhat subjective.

Visually what we are looking for is the data that hugs the line tightly or is sitting close to the line; this would give us confidence in our assumption that the residuals are normally distributed. Now, it is highly unlikely that the data will perfectly hug the line, so our observation and conclusion are very subjective.

## Checking the Mean of the Residuals Equals 0

Additionally, we need to check another assumption, that the mean of the residuals is equal to zero. A mean value close to zero is a good thing and we can proceed to the next step. On a side note, it is not uncommon to get a mean value that is not exactly zero; which happens due to rounding errors. However, if the mean value is very close to zero, we can confidently use it. In the example below, you may see that the mean value is not exactly zero.



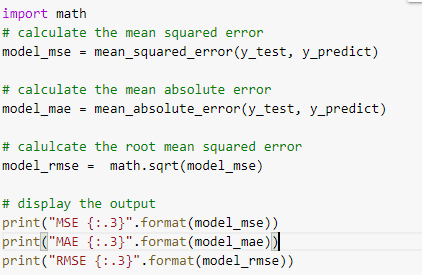


## Measures of Error

We next examine that how well our data fit the model, so we take the y\_predictions and compare them to our y\_actuals which will be our residuals. From here we can calculate a few metrics to help quantify how well our model fits the data. Here are a few popular metrics:

* **Mean Absolute Error (MAE):** MAE Is the mean of the absolute value of the errors. This gives us an estimate of magnitude but no sense of direction (too high or too low).
* **Mean Squared Error (MSE):** MSE Is the mean of the squared errors. MSE is more popular than MAE in such calculations, because MSE "*punishes*" more significant errors.
* **Root Mean Squared Error (RMSE):** RMSE Is the square root of the mean of the squared errors. RMSE is even more favored in such calculations, because it allows us to interpret the output in y-units.

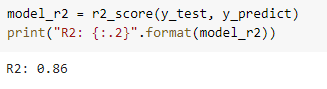
Fortunately, both **sklearn** and **statsmodel** contain functions, that can calculate these metrics. The examples below were calculated using the *sklearn* library and the *math* library.





## R-Squared

The R-Squared metric provides us a way to measure the goodness of fit or, in other words, how well our data fits the model. A higher value of the R-Squared metric means a better fit for our model. However, one limitation of this calculation is that the value of R-Square increases as the number of features increase in our model. This implies that if we keep adding variables including poor choices, the R-Squared will go further up! **A more popular metric is the adjusted R-Square which penalizes more complex models, or in other words, the models with more exploratory variables.** In the example below, the regular R-Squared value has been calculated; however, the statsmodel summary calculates the adjusted R-Squared below.



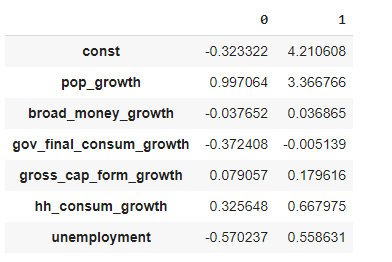
## Confidence Intervals

Let us look at our confidence intervals. Keep in mind that by default, the confidence intervals are calculated using 95% intervals. We interpret confidence intervals by saying that if we sample our target population a 100 times, **approximately 95 of those confidence intervals would contain the "true" coefficient.**

Why do we provide a confidence range? Well, this is because we only have a sample of the population, and not the entire population to collect data from. Because of this limitation, that the "true" coefficient may exist in the interval below or it may not exist although we are not sure about that. We provide some uncertainty by providing a range, usually 95%, where the coefficient is probably in.

* Want a narrower range? **Decrease your confidence**.
* Want a wider range? **Increase your confidence**.



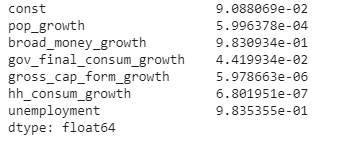


## Hypothesis Testing

With hypothesis testing, we try to determine the statistical significance of the coefficient estimates. This test is outlined as the following:

* **Null Hypothesis:** There is no significant relationship between the exploratory and the explanatory variables.
* **Alternative Hypothesis:** There is a significant relationship between the exploratory and the explanatory variables.
* If we **reject the null**, we imply that there is a significant relationship between exploratory and explanatory variables, but the coefficients do not equal 0.
* If we **fail to reject the *null***, it means that there is no relationship, and the coefficients do equal 0.



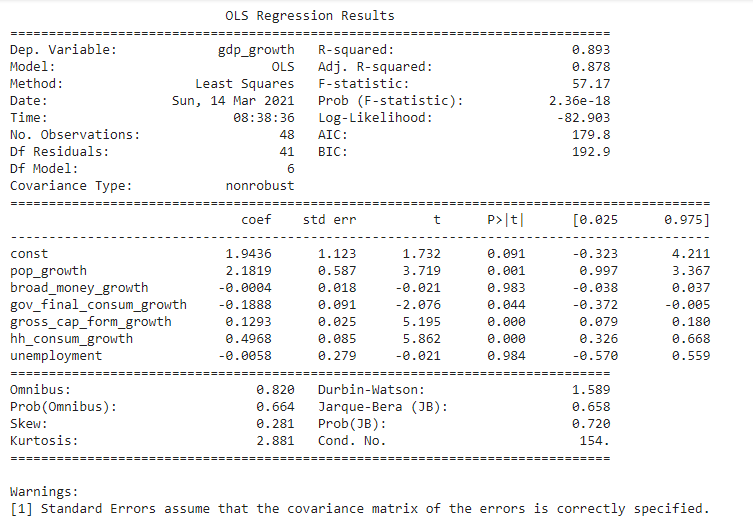


This is difficult to explain, but we have a few insignificant coefficients in such analysis. The first of them is the constant itself, so technically this should be dropped. However, we will see that once we remove the irrelevant variables, the intercept becomes significant. **If it is still not significant, we start the intercept at 0 and assume that the cumulative effect of X on Y begins from the origin (0, 0).**  Along with the constant, we have *unemployment* and *broad\_money\_growth* both come out as insignificant.

## Create a Summary of the Model Output

Let us create a summary of some of our keep metrics; *sklearn* does not have a good way of creating this output, so we calculate all the parameters ourselves. Let us avoid this calculation and use the statsmodel.api library. With this library, we can create the same model we created above, but we also leverage the *summary()* method to create an output. Some of the metrics might differ slightly, but they generally should be the same.





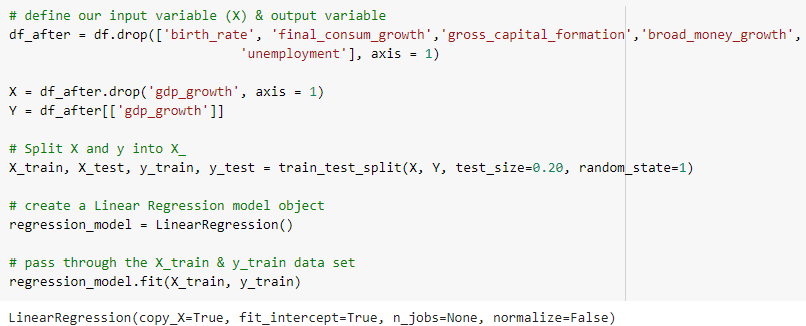
The first thing to notice is that the p-values are now easier to read and we can now remove the coefficients with p-values greater than 0.05. We also have the 95% confidence interval (described up above), coefficient estimates (described up above), standard errors, and t-values.

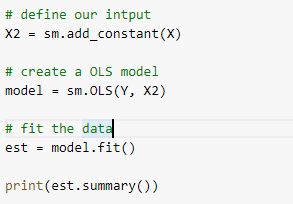
The other metric that stands out is the Adjusted R-Squared value which is 0.878, lower than the R-Squared value. This makes sense since we were probably docked for the complexity of our model. However, an R-Squared over 0.878 is still very strong.

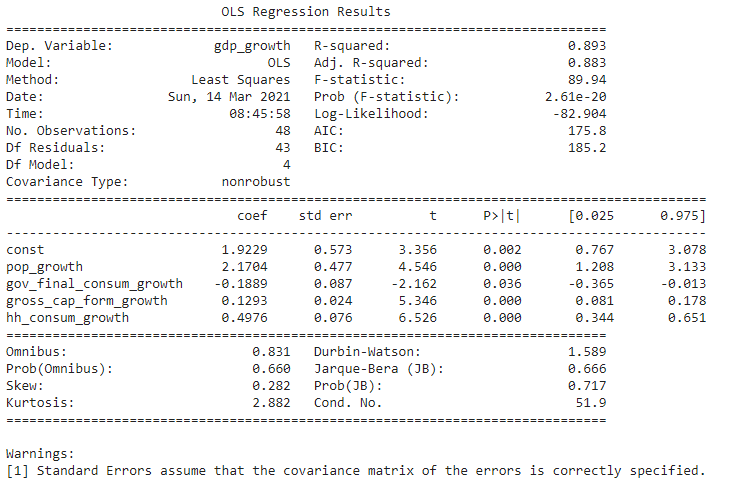
The only additional metrics we describe here is the t-value which is the coefficient divided by the standard error. The higher the t-value, the more evidence we have to reject the null hypothesis. Also, remember that the standard error is the approximate standard deviation of a statistical sample population.

## Remove the Insignificant Variables

Now that we have identified the insignificant variables, we should remove them from the model and refit the data to see what we get. The steps are the same. The only thing I have changed is the removal of additional columns from the data frame.



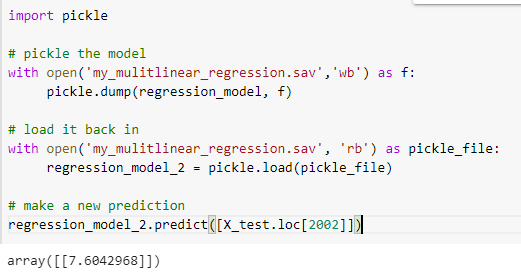




**Looking at the output, we find that all the independent variables are significant and the constant is significant.** We could rerun our test for autocorrelation and, but this will take you to the same conclusions we found above, so I have decided to leave that from the chapter. At this point, we can interpret the formula and begin making predictions. Looking at the coefficients, we can say that the *pop\_growth, gross\_cap\_from\_growth,* and *hh\_consum\_growth* all have a positive effect on GDP growth. Additionally, we can say that the gov\_final\_consum\_growth has a negative effect on GDP growth. This is a slightly surprising finding, but we will have to find its reason.

## Save the Model for Future Use

We will probably use this model in the future, so let us save our work for future use. Saving the model can be achieved by storing our model in a pickle, which is storing a *python* object as a character stream in a file, which can be later reloaded for use.



## Summary

Regression models describe the relationship between variables by fitting a line to the observed dataset. Generally, the linear regression models use a straight line. Regression allows us to estimate how a dependent variable changes with the change in the independent variables. Simple linear regression is used to estimate the relationship between two quantitative variables, whereas multiple linear regression is utilized to estimate the relationship between two or more quantitative and qualitative variables.

## Assessment

## Choose the appropriate option

1. **Which one of the following are regression tasks?**
   1. Predict the age of a person
   2. Predict the country from where the person comes from
   3. Predict whether the price of petroleum will increase tomorrow
   4. Predict whether a document is related to science
2. **Which of the following plots is used for normality test?**
   1. Scatter plot
   2. Bar plot
   3. qqplot
   4. None of these
3. **Which of the following tests is used for heteroscedasticity?**
   1. AD
   2. Ljung-Box
   3. Breusch-Pagan
   4. All of the above
4. **Which of the following tests is used for *autocorrelation*?**
   1. AD
   2. Ljung-Box
   3. Breusch-Pagan
   4. White test
5. **VIF > 10 is said to be:**
   1. No Multicollinearity
   2. Less Multicollinearity
   3. High Multicollinearity
   4. None of the above

## Fill in the spaces with appropriate answers

1. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ diagrams are graphs of the data that are helpful in displaying the relationships between variables.
2. The SSR is sometimes referred to as the variability in Y, as explained by \_\_\_\_\_\_\_\_\_\_\_\_\_\_.
3. If the adjusted R2 \_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_ when a new variable is added, it would be an indication that the variable should not remain in the model.
4. Regression analysis is sometimes called \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.
5. Complete the following equation: SST = SSR +

## True or False

1. Simple Linear regression is built among more than two variables.
   1. True
   2. False
2. Multiple Linear regression is built between two or more than two variables:
   1. True
   2. False
3. Autocorrelation is having relationship between observation:
   1. True
   2. False
4. We drop the variable from the model if the p-value is insignificant:
   1. True
   2. False
5. Difference between actual and predicted values is called errors (or) residuals:
   1. True
   2. False

## Programming Assignment

Using the data in the below URL,

<https://www.kaggle.com/ruiromanini/mtcars>

1. Build the multiple linear regression model and predict mpg.
2. Validate the assumptions for multiple linear regression.
3. Evaluate the model metrics on the test set.

**Solutions:** Refer to page 270

## Solutions for Assessment

## Choose the appropriate options

1. A
2. C
3. C
4. B
5. C

## Fill in the spaces with appropriate answers

1. Scatter
2. The regression equation
3. Decreases
4. Least-Squares Regression
5. SSE

## True or False

1. False
2. True
3. True
4. True
5. True

# CHAPTER 7: Regression Analysis – Part II

## Theory

In the previous chapter, we studied the regression analysis and its assumptions. In this chapter we will investigate the Cost Function Optimization Algorithm, Categorical Features in a Regression model, Feature Scaling, Features Selection, Regularization, and Polynomial Regression.



In the previous chapter, we discussed a few cost functions such as,

* Mean Absolute Error
* Mean Square Error
* Root Mean Squared Error

## Cost function Optimization Algorithm

The cost function optimization algorithm attempts to find the optimal values of the model parameters by finding the global minima of cost functions. One of the most important cost functions used in the regression analysis is the Gradient Descent.

## What is Gradient Descent?

Let’s assume you are playing a game where the players are at the top of a mountain, and they are asked to reach the lowest point of the mountain. Additionally, they are blindfolded. Thus, what approach do you think would help you reach the lowest point?

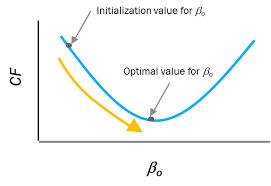
Take a moment to think about this problem before you read further.

The best approach is to observe the ground and find the point where the land descends. From that position, take a step in the descending direction and iterate this process until you reach the lowest point.



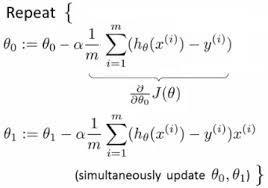
To find the local minimum of a function using gradient descent, we must take steps, which are proportional to the negative value of the gradient (as we move away from the gradient) function at the current point. If we take steps, which are proportional to the positive value of the gradient (moving towards the gradient), we will reach a local maximum of the function, and this procedure is called Gradient Ascent.

The algorithm of Gradient Descent was originally proposed by CAUCHY in 1847. It is also known as the steepest descent.

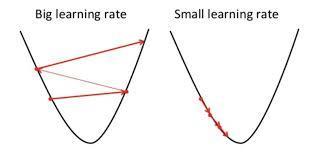


The goal of the gradient descent algorithm is to minimize the given function (*e.g.*, the cost function). To achieve this goal, it iteratively performs two steps:

1. Compute the gradient (slope) by determining the first-order derivative of the function at that point.
2. Subsequently, make a step (move) in the direction opposite to the gradient. The opposite direction of slope increases from the current point by alpha times the gradient at that point.



Where *𝜃*0 is the ‘intercept’, *𝜃*1 is the slope, *𝛼* is the learning rate, m is the total number of observations, and the sign after the ∑ sign is the loss. Google Tensor board recommends a learning rate between 0.00001 and 10. Generally, a smaller learning rate is recommended to avoid overshooting while creating a model.



We can visualize the gradient descent optimization algorithm by clicking on the following link.

[https://miro.medium.com/proxy/0\*D7zG46WrdKx54pbU.gif](https://miro.medium.com/proxy/0*D7zG46WrdKx54pbU.gif)



In regression analysis sometimes, we come across variables that are qualitative or on a nominal scale, in nature, such as gender, color, nationality, geographical region, etc.

In such cases, we transform the nominal variables into many dichotomous variables (like “Yes” or “No”) which are called “dummy variables”. Dummy Variables assume the values of 0 and 1. Therefore, such variables are essentially a device to classify data into the mutually exclusive categories, such as male or female.

**Example:**

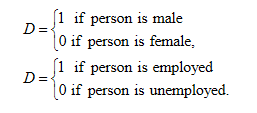
The manager of a small sales force wants to know whether the average monthly salaries are different for the male and female employees in the sales force. He obtains data on monthly salary and experience (in months) for each of the 9 employees as presented in the table below.

| **Employee** | **Salary** | **Gender** | **Experience** |
| --- | --- | --- | --- |
| 1 | 7.5 | Male | 6 |
| 2 | 8.6 | Male | 10 |
| 3 | 9.1 | Male | 12 |
| 4 | 10.3 | Male | 18 |
| 5 | 13 | Male | 30 |
| 6 | 6.2 | Female | 5 |
| 7 | 8.7 | Female | 13 |
| 8 | 9.4 | Female | 15 |
| 9 | 9.8 | Female | 21 |

Now, creating a dummy variable for gender, we have:

| **Employee** | **Salary** | **Gender** | **Experience** |
| --- | --- | --- | --- |
| 1 | 7.5 | 0 | 6 |
| 2 | 8.6 | 0 | 10 |
| 3 | 9.1 | 0 | 12 |
| 4 | 10.3 | 0 | 18 |
| 5 | 13 | 0 | 30 |
| 6 | 6.2 | 1 | 5 |
| 7 | 8.7 | 1 | 13 |
| 8 | 9.4 | 1 | 15 |
| 9 | 9.8 | 1 | 21 |

Usually, the dummy variables take on the values 0 and 1 in order to identify the mutually exclusive classes of the explanatory variables. For example,



It is also not mandatory to choose only 1 and 0 to denote the category. In fact, any distinct value of D can serve the purpose. The choices of 1 and 0 are preferred as they make the calculations simple, help in an easy interpretation of the values and usually turn out to be a satisfactory choice.

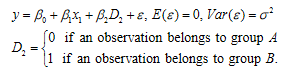
In a given regression model, the qualitative and quantitative variables can also occur together*, i.e*., some variables are qualitative, and others are quantitative. Accordingly, when all explanatory variables are

* In the case of quantitative, the model is called a regression model.
* In the case of qualitative, the model is called an analysis of variance model.
* Quantitative and Qualitative both, then the model is called an analysis of covariance (ANOVA) model.

Such models can be tackled within the framework of the regression analysis. Moreover, the usual tools of regression analysis can be used in the case of dummy variables.

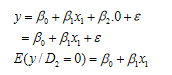
**Example:**

Consider the following model with X1 as quantitative and D2 as a dummy variable



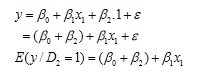
The interpretation of the results is essential. We proceed as follows:

If D2 = 0, then



Where the relationship with intercept ®0 and slope ®1 is a straight line.

If D2 = 1, then



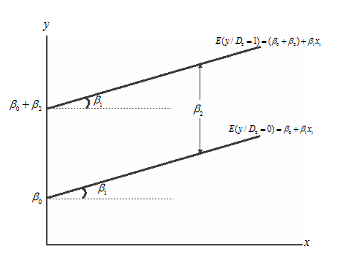
Where the relationship with intercept (®0 + ®2) and slope ®1 is a straight line.

The quantities E(y / D2 = 0) and E(y / D1 = 0) are the average responses, when an observation belongs to the groups A and B, respectively. Thus

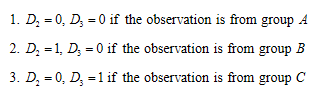


Which has an interception as the difference between the average values of Y with D2 = 0 and D2 = 1.

Graphically, it looks like the following figure, which describes the two parallel regression lines with the same variances ⌠2.



If there are three explanatory variables in the model with two dummy variables D2, and D3 then they will describe three levels, e.g., groups A, B, and C. In such case, the levels of dummy variables will be as follows:



The concerned regression model is given as:



In general, if a qualitative variable has m levels, then (m – 1) dummy variables are required for such analysis, and each of these variables takes a value of 0 or 1.

Consider the following examples to understand how to define and handle the dummy variables.

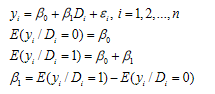
**Example:**

Suppose the symbol ‘’Y’’ denotes the monthly salary of a person and ‘’D’’ denotes whether the person is a graduate or non-graduate.

The model is



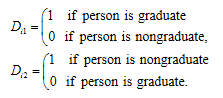
With n number of observations, the model is



Here,



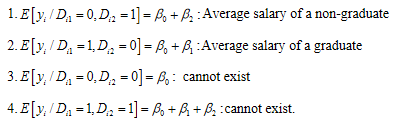
Now consider the same model with two dummy variables, as defined in the following way:



The model with n number of observations is



Then we have



Notice that in this case



Which is an exact constraint and indicates the contradiction in the analysis as follows:



The analysis shows that multicollinearity is present in such cases. Hence, the rank of the matrix of explanatory variables falls short by 1. Consequently, ®0, ®1, and ®2 are indeterminate, and the least-squares method breaks down. So the proposition of introducing two dummy variables is useful, but it can lead to serious consequences. This can also be termed as the **dummy variable trap**.

However, if the intercept term is ignored, then the model becomes



Then



Therefore, when the intercept term is dropped, the ®1 and ®2 have proper interpretations as average salaries of the graduate and non-graduate persons, respectively.

Now the parameters can be estimated using the ordinary least squares principle, and the standard procedures for drawing inferences can be utilized.

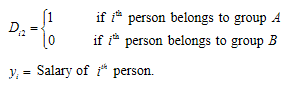
**Rule of thumb:** When the explanatory variable leads to the ' ‘’m’’ mutually exclusive categories classification, then use (m-1) dummy variables for its representation. Alternatively, use the ‘’m’’ dummy variables but drop the intercept term.

**Interaction term:**

Suppose a model has two explanatory variables – one is quantitative and the other is a dummy. Suppose both variables interact and we add an explanatory variable as the interaction of them to the model.



To interpret the model parameters, we can proceed as follows: Suppose that the dummy variables are given by,



Then,



This is a straight line with intercept ®0 and slope ®1. Next, we get



This is a straight line with the intercept term (®0 + ®2) and slope (®1 + ®3).

Hence, we obtain this model,



Which has different slopes and intercept terms.

Thus, ®2 reflects the change in the intercepts associated with the change in the group of persons *i.e*., when the group changes from A to B.

®3 reflects the change in the slope associated with the change in the group of people, i.e., when the group changes from A to B.

Fitting of the model results in,



Which is equivalent to fitting two separate regression models corresponding to Di2 = 1 and Di2 = 0, i.e.



And



respectively.

In this way, using a dummy variable makes the testing of hypotheses more convenient. For example, if we want to test whether the two regression models are identical, the test of hypothesis involves testing of



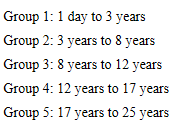
Here, the acceptance of H0 indicates that only a single model is necessary to explain the relationship.

In another example, if our objective is to test whether the two models differ with respect to intercepts only and have the same slopes, then we will test this hypothesis using



## Dummy variables versus quantitative explanatory variable

The quantitative explanatory variable can be converted into dummy variables. For example, if the ages of people in certain groups are as follows:



Subsequently, the variable ‘*age’* can be represented by four different dummy variables.

Since it is difficult to collect the data on individual ages, this method helps in an easy collection of data. A disadvantage of this method is the loss of information to some extent. For example, if the ages are in years 2, 3, 4, 5, 6, 7, then the dummy variable is defined as



Then these values become 0, 0, 0, 1, 1, and 1. Now looking at the value 1, one cannot determine if it corresponds to age 5, 6, or 7 years.

Moreover, If a quantitative explanatory variable is grouped into ‘’m’’ categories, then (m – 1) parameters are required for analysis. However, if the original variable is used as such, then only one parameter is required for analysis.

Treating a quantitative variable as a qualitative variable increases the complexity of the model, and reduced the degrees of freedom for error. These alterations can affect the inferences if the dataset is small. Contrary, in a large dataset, such an effect may be small.

It is evident that the usage of dummy variables does not require any assumption about the function form of the relationship between study and explanatory variables.



## Why do we need to perform feature scaling?

The real-world dataset has multiple features spanning varying degrees of magnitude, ranges, and units. This is a significant obstacle since the regression analysis is highly sensitive to these features. Plotting features on a similar scale can help the gradient descent converge faster towards the minima.

## Feature scaling techniques

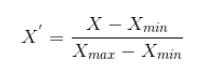
Following are the two important techniques in feature scaling.

* Normalization
* Standardization

## Normalization

Normalization is a scaling technique in which values are shifted and rescaled, so that they end up in a range from values 0 and 1. This technique is also known as Min-Max scaling.

Here is the formula for normalization,



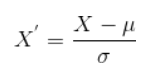
In this formula, *Xmax* and *Xmin* are the maximum and the minimum values of the features, respectively.

* When the value of X is the minimum value in the column, the numerator will be 0, and hence X’ = 0.
* On the other hand, when the value of X is the maximum value in the column, the numerator is equal to the denominator and thus the value of X’ = 1.
* If the value of X is between the minimum and maximum value, then the value of X’ can be any number between 0 and 1.

## Standardization

Standardization is another scaling technique, in which the values are centered around the mean value with a unit standard deviation. This implies that the mean value of the attribute becomes zero and the resultant distribution has a unit standard deviation.

Here is the formula for standardization:



Where µ is the mean and ⌠2 is the standard deviation of the feature values. Note that in this case, the values are not restricted to a particular range.



Regression Analysis works on a simple rule – If you put garbage in, you will get garbage out. There is no way to filter out good data in regression analysis. By garbage here, I mean noise in the dataset.

This notion becomes even more important when the number of features are very large. You do not need to use every feature at your disposal for creating an algorithm. Instead, you can assist the algorithm by feeding in only the important and relevant features. I have myself witnessed that feeding subsets of features to an algorithm yields better results than while feeding a complete set of features to the same algorithm.

This strategy has its uses beyond competition and in industrial applications as well. Applying this strategy not only reduces the time for evaluation and training, but also ensures that you have fewer things to worry about!

Some common feature selection methods are:

* Backward Elimination
* Forward Selection
* Stepwise Regression

## Backward Elimination

This is the simplest of all variable selection procedures and can be easily implemented with the following steps

1. Start with considering all the predictors in a given model.
2. Remove the predictor with highest p-value, which is greater than αcritical.
3. Refit the model and repeat step 2.
4. Stop when all p-values are less than αcritical.

The αcritical does not have to be 5%. If achieving a good prediction performance is the goal, then a 15-20% cut-off may work best, although methods which are designed more directly for optimal prediction should be preferred.

## Forward Selection

This procedure just reverses the backward method.

1. Start with no variable in the model.
2. For all predictors which are not in the model, check their p-value if they are added to the model. Choose the one with the lowest p-value less than αcritical.
3. Continue until no new predictors can be added.

## Stepwise Regression

This procedure is a combination of backward elimination and forward selection. It addresses the situation where variables are added or removed early in the process and we want to change our mind about their possible use later. At each stage, a variable may be added or removed and there are several procedures to perform these steps. Stepwise procedures are relatively less computationally but they do have some drawbacks.

1. Because of the “one-at-a-time” nature of adding/dropping variables, it is possible to miss the “optimal” model.
2. The p-values used should not be treated too literally. This is because multiple testing is being implied simultaneously, which can reduce the validity of the p values. The removal of less significant predictors tends to increase the significance of the remaining predictors. This effect may result in overstating the importance of the remaining predictors.
3. The procedures are not directly linked to the final objectives of prediction or explanation and hence may not really help solve the problem of interest. With any variable selection method, it is imperative to keep in mind that the model selection cannot be divorced from the underlying purpose of investigation. Variable selection tends to amplify the statistical significance of the variables that stay in the model. The dropped variables can still be correlated with the response. It would be wrong to say these variables are unrelated to the response. Rather, they provide no additional explanatory effect beyond the variables already included in the model.
4. Stepwise variable selection tends to pick models that are smaller than desirable for the purpose of prediction. To give a simple example, consider the simple regression with just one predictor variable. Suppose that the slope for this predictor is not quite statistically significant. We might not have enough evidence to say that it is related to ‘’y’’ but it still might be better to use it for predictive purposes.

## Underfitting and Overfitting

## Underfitting

A statistical model is said to have Underfitting when it cannot capture the underlying trend of the data. For example, what if I send a 3rd grade kid to a Differential Calculus Class; the kid is only familiar with the basic arithmetic operations. On a similar analogy, If the data contains too much information that the model cannot take, the model is likely going to underfit.

This situation usually happens if we have less data to train our model, but quite a large number of features, or when we try to build a linear model with non-linear data. In such cases, the rules of the regression model are too comfortable and flexible to be applied on such minimal data; and, therefore the model will probably make a lot of wrong predictions.

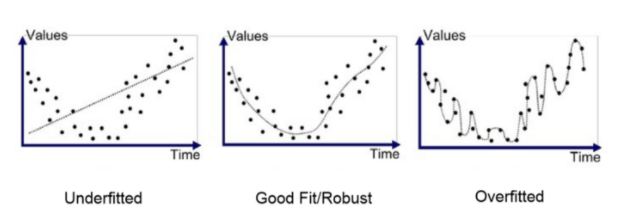
Specifically, the underfitting occurs if the model or algorithm shows a high bias. Underfitting is often a result of an excessively simple model.

If a model does not perform well including on the training data, then the model is most likely under fitted.

## Overfitting

Overfitting occurs when a statistical (or regression) model captures the noise of the data. Intuitively, overfitting occurs when the model or the algorithm fits the data too well. Specifically, overfitting occurs if the model or algorithm exhibits low bias but high variance. Overfitting is often a result of an excessively-complicated model applied to an uncomplicated dataset.

If a model performs very well on the training data but fails to perform similarly on the test data, then this model is most likely overfitted.



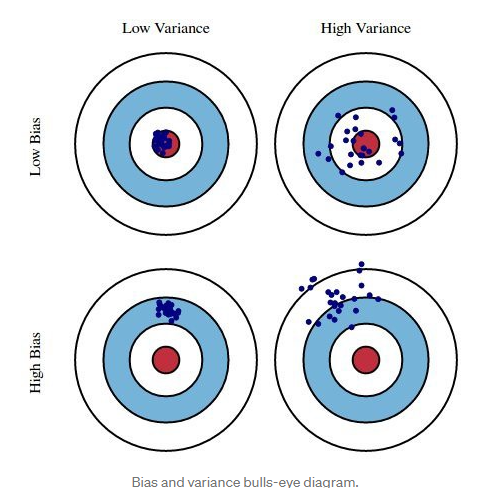
Essentially, we always need to develop a model which performs well on both the training and test datasets, and we can term it as a *generalized model*.

## Bias and Variance Tradeoff

An important theoretical result of statistics (and regression analysis) is the fact that a model’s generalization error can be expressed as a sum of three very different errors:

* **Bias:** This part of the generalization error is due to wrong assumptions. For example, assuming that the data is linear, when it is actually quadratic. A high-bias model is most likely to underfit the training data.
* **Variance:** This part is due to the model’s excessive sensitivity to small variations in the training data. A model with multiple degrees of freedom (Such as a high-degree polynomial model) is likely to have high variance, and thus to overfit the training data.
* Irreducible error: This part of generalization error is due to the noisiness of the data itself. The only way to reduce this part of error is to clean the data (*e.g.,* fix the data sources, such as broken sensors, or detect and remove the outliers).

Biase and variance are inversely related so increasing the bias decreases variance, and increasing the variance decreases bias. A model that exhibits **low variance and high bias will underfit** the target, while a model with **high variance and low bias will overfit** the target. Our primary objective is to reach a model with low bias and low variance.

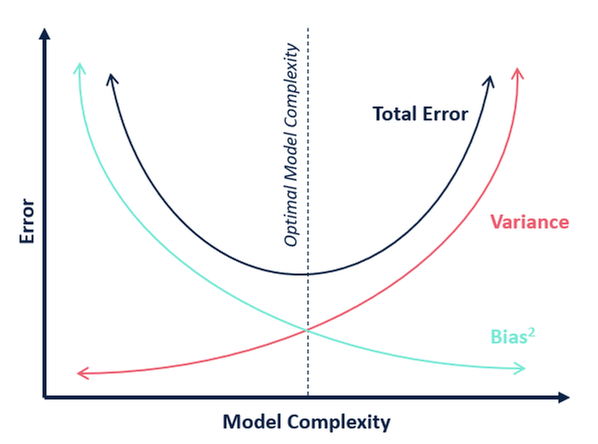


In the above diagram, the center of the circles *i.e.* the bull’s eye is the target that the model tries to predict correctly. As we move away from the bulls-eye, the model starts to make more and more wrong predictions.

A model with low bias and high variance predicts the locations of points around the center, but far away from each other. A model with high bias and low variance is far away from the bull’s eye, but since the variance is low, the predicted points are closer to each other.

Therefore, the main challenge is to find the right balance between the bias and variance of the model.

In terms of the model complexity, we can use the following diagram to decide on the optimal complexity of our model.



As evident from the diagram, as the complexity of the model increases, the bias error comes down. However, the complexity of the model keeps increasing, the variance error also starts to increase. The intercept of bias and variance is said to be the optimal model complexity.

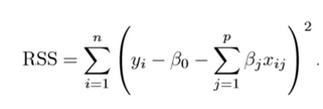
## Regularization

When we use the regression models to train the data, there is a good chance that the model will overfit the given training dataset. Regularization helps sort this overfitting problem by restricting the degrees of freedom of a given equation *i.e.* simply reducing the number of degrees of a polynomial function by reducing the corresponding coefficients.

In a linear regression, we do not want huge coefficients as a small change in the coefficient can make a large difference for the independent variable (Y). So, regularization constraints the coefficients of such features to avoid the problem of overfitting. Simple linear regression is given as:



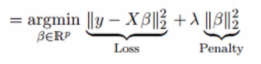
Using the OLS method, we try to minimize the cost function given as:



To regularize the model, a shrinkage penalty is added to the cost function. Let’s now investigate the different types of regularizations in regression:

* **Ridge Regression (L2 Form):** Ridge regression is a technique used when the data suffers from multicollinearity (independent variables are highly correlated). In multicollinearity, even though the ordinary least squares estimates (OLS) are unbiased, their variances are large, which results in a deviation of the observed value far from the true value. By adding a degree of bias to the regression estimates, ridge regression reduces the standard errors.

Ridge regression solves the multicollinearity problem through shrinking parameter λ. Look at the equation below



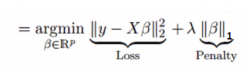
In this equation, we have two components. First one is the least square term and the second one is the lambda of the summation of β2 (beta – square) where β is the coefficient. This is added to least square term in order to shrink the parameter to have a very low variance.

**Important Points:**

* + The assumptions of the ridge regression is the same as OLS regression except that the normality is not assumed
  + Applying ridge regression shrinks the value of coefficients but the values do not reach zero, which suggests that there is no feature selection.
* **LASSO (L1 form):** The LASSO (Least Absolute Shrinkage and Selection Operator) regression penalizes the model based on the sum of magnitude of the coefficients. The regularization term is given by the following equation:



Where, λ is the shrinkage factor and hence the formula for loss after regularization is:



Similar to Ridge regression, LASSO also penalized the absolute size of the regression coefficients. Additionally, it reduces the variability and improves the accuracy of linear regression models. Look at the equation above: LASSO regression differs from ridge regression because instead of squares, it uses absolute values in the penalty function. This leads to penalizing values which causes some of the parameter estimates to become exactly zero. Larger the penalty applied, further the estimates get shrunk towards absolute zero, which results in variable selection out of a given n number of variables.

**Important Points:**

* The assumption of LASSO regression is the same as least square regression except that the normality is not to be assumed.
* Application of LASSO regression shrinks the coefficients to zero (exactly zero), which certainly helps in the feature selection.
* If a group of predictors are highly correlated, LASSO regression picks only one of them and shrinks the other predictors to zero.
* **ElasticNet Regression:** ElasticNet is a hybrid of LASSO and ridge regression techniques. It is trained with L1 and L2 prior to regularization. ElasticNet is useful when there are multiple correlated features. LASSO is likely to pick one of these at random, while ElasticNet is likely to pick both.



A practical advantage of trading-off between LASSO and Ridge is that it allows ElasticNet to inherit some of Ridge’s stability under rotation.

**Important Points:**

* It encourages the group effect in case of highly correlated variables.
* There are no limitations on the number of selected variables.
* It can suffer from double shrinkage.

## Polynomial Regression

For understanding Polynomial Regression, let’s first understand what a polynomial is. Merriam-webster defines a polynomial as: “A mathematical expression of one or more algebraic terms, each of which consists of a constant multiplied by one or more variables raised to a non-negative integral power (such as a + bx + cx2)”. Simply said, *poly* means many. Therefore, a polynomial is an aggregation of many monomials (or Variables). A simple polynomial equation can be written as:





In the equation,



The maximum power of ‘X’ is called the *degree* of the polynomial equation. For example, if the degree is 1, the equation becomes



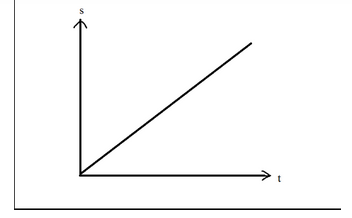
This is a simple linear equation. If the degree is 2, then the polynomial equation becomes:



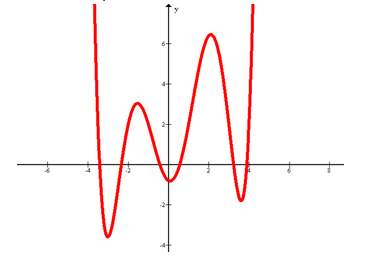
This is known as a quadratic equation and so on.

## When to use Polynomial Regression?

Several times in our data analysis, we may face a requirement where we have to do a regression analysis. However, when we plot a graph between dependent and independent variables, it does not turn out to be a linear one. A linear graph typically looks like:

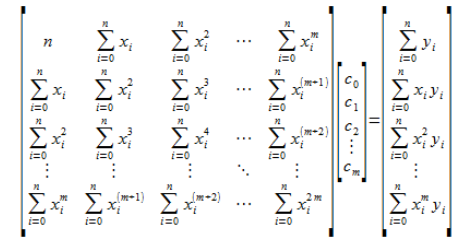


What if the relationship looks like this?



It means that the relationship between X and Y cannot be described linearly. In such scenarios, we may use Polynomial Regression.

We can generalize the matrix obtained for linear regression equation of n number of coefficients (in y = mx + b, m and b are the coefficients) as follows:



Where ' ‘’m’’ is the degree (maximum power of ‘’x’’) of the polynomial and ‘’n’’ is the number of observation points. The above matrix results in the general formula for polynomial regression. Earlier, we were able to visualize the calculation of minima because the graph was presented in the three dimensions. But as there are ‘’n’’ number of coefficients, it is not possible to create an (n+1) dimension graph here.

## AIM

The aim of the following lab exercises is to perform various exercises by writing the corresponding python codes, so that we can get hands-on practice of the descriptive statistics.

The labs for this chapter include the following exercises.

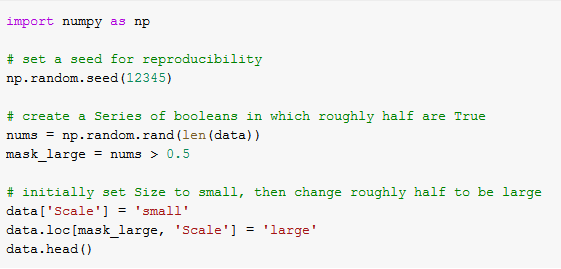
* Handling Categorical Predictors
* Regularization
* Polynomial Regression

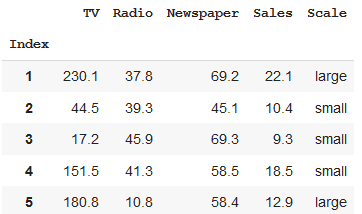
For these exercises, we will use python3 and jupyter notebook IDE.

## Task 1): Handling Categorical Predictors with Two Categories

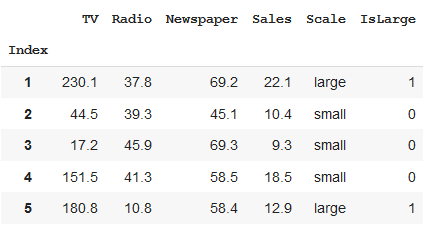
Till now, all the predictors have been numeric; what if one of the predictors is categorical?

We will use advertisement data for this task. We’ll create a new feature called Scale, and will randomly assign observations as small or large.



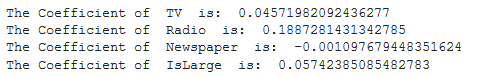


For the *scikit-learn* library, all the data must be represented numerically. If a feature has only two categories, we can simply create a dummy variable which represents the categories as a combination of binary values.



Let us redo the multiple linear regression problems and include the “*IsLarge*” predictor as follows:



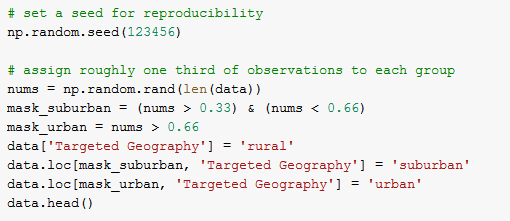


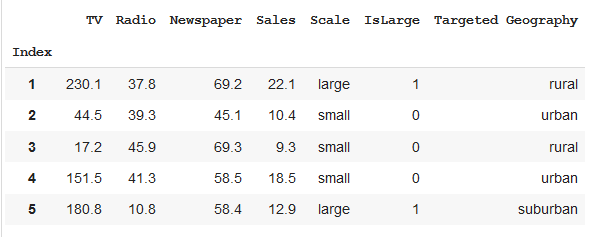
How do we interpret the coefficient for IsLarge? For a given TV/Radio/Newspaper ad expenditure, if the average sales increase by 57.42 widgets, we can consider it as a large market.

What if the 0/1 encoding is reversed? The value of the coefficient will still be the same, however the sign will change from positive to negative.

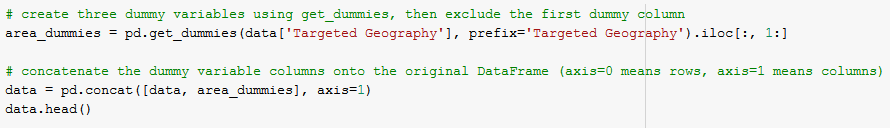
## Task 2) Handling Categorical Variables with More Than Two Categories

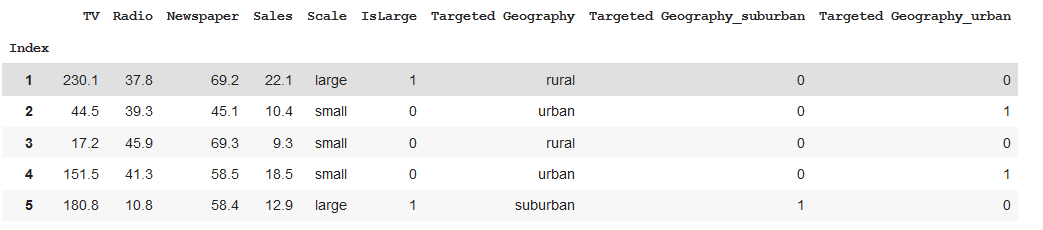
Let us create a new column called the **Targeted Geography**, and randomly assign observations to be **rural, suburban or urban**.





Here, we need to represent the ‘Targeted Geography’ column numerically. But mapping urban=0, suburban=1, and rural=2 will mean that the rural area is two times more suburban which is not the case. Hence, we will create another dummy variable as follows:





What does the encoding say?

* Rural is encoded as Targeted Geography\_suburban=0 andTargeted Geography\_urban=0
* Suburban is encoded as Targeted Geography\_suburban=1 and Targeted Geography\_urban=0
* Urban is encoded as Targeted Geography\_suburban=0 and Targeted Geography\_urban=1

Now the question is: Why have we used two dummy columns instead of three?

Because by using only two dummy columns, we can capture the information of all the 3 columns. For example, if the values for Targeted Geography\_urban and Targeted Geography\_rural are 0, we can infer that the data belongs to Targeted Geography\_suburban.

This is called handling *the dummy variable trap*. If there are ' ‘’m’’ number of dummy variable columns, then the same information can be conveyed by the ' ‘’m-1’’ columns. Let’s include the two new dummy variables in the model.





How do we interpret the coefficients?

* If all other columns are constant, then the suburban geography is associated with an average **decrease** of 106.56 widgets in sales for $1000 spent.
* If $1000 is spent in an urban geography, it accounts for an average **increase** in Sales of 268.13 widgets.

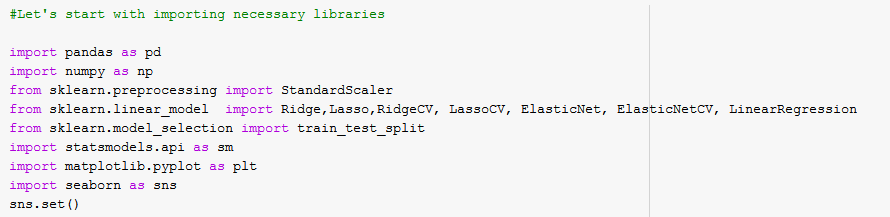
A final note about the dummy encoding: If we have categories that can be ranked in an order (*i.e.,* worst, bad, good, better, and best), we can potentially represent them numerically as (1, 2, 3, 4, and 5 respectively) using a single dummy column.

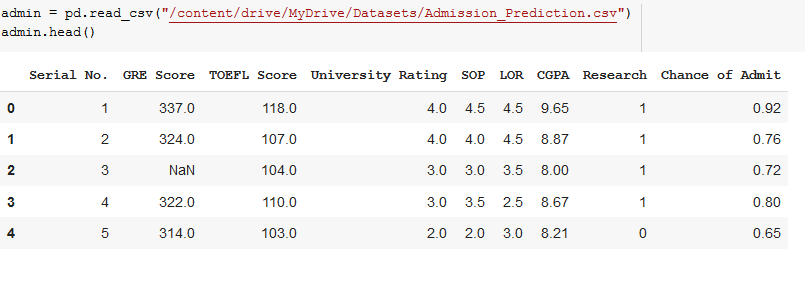
## Task 3) Regularization the algorithms

**Problem statement**

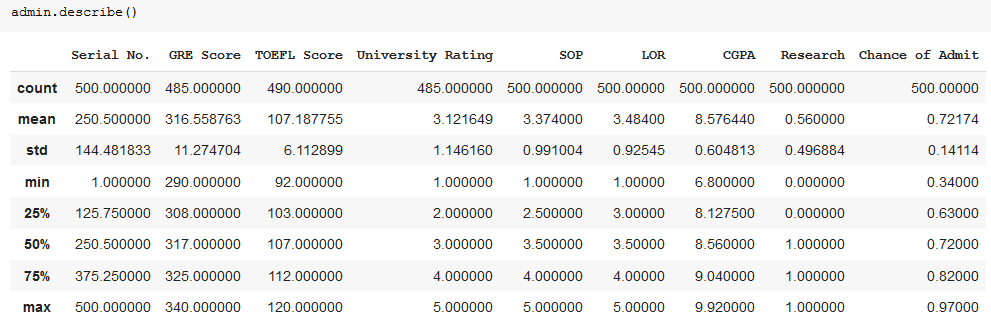
The dataset contains several important parameters to consider during the application for Masters Programs. This dataset was built with the purpose of helping students in shortlisting universities with their profiles. The predicted output gives them a fair idea about their chances of admission into a particular university.

**Importing necessary packages**

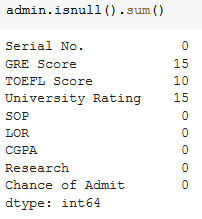




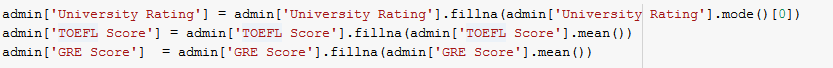
**Descriptive statistics**



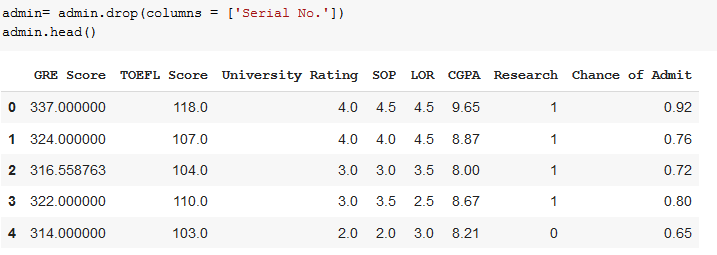
**Checking the missing values**



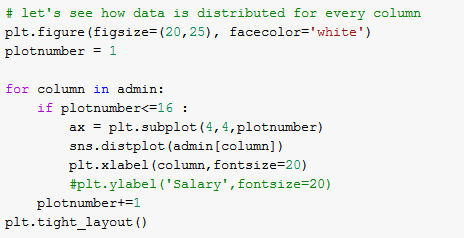
There are a few missing values in GRE Score, TOEFEL Score, and the University Rating variables. Let us impute the missing values as follows:

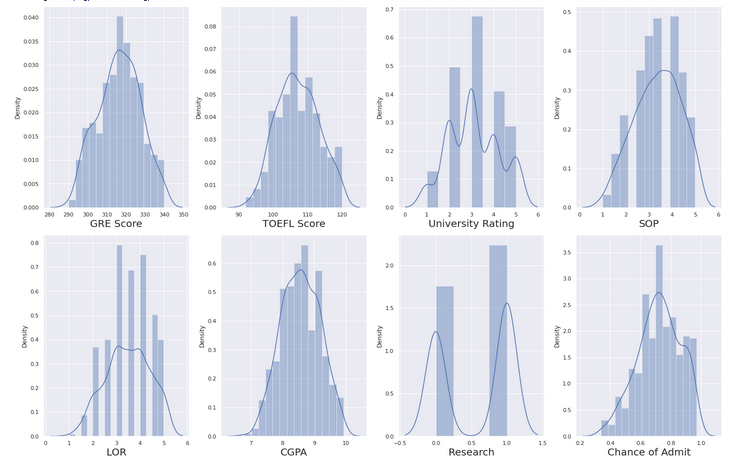


Now the data looks good and there are no missing values. Also, the first column includes the serial numbers, which are not needed for statistical analysis and can be omitted.



Let’s visualize the data and analyze the relationship between independent and dependent variables.

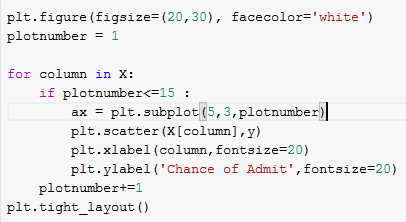


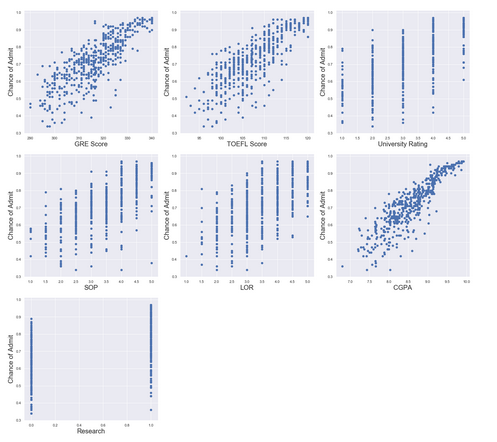


The data distribution looks close to normal without showing any skewness. Great so let’s go ahead!

Let us observe the relationship between the independent and dependent variables.



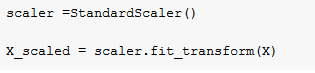




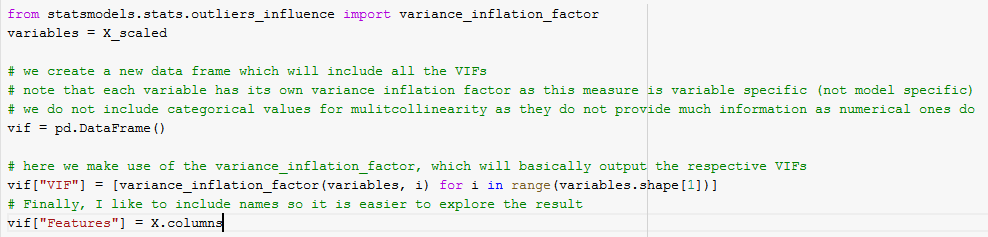
The relationship between the dependent and independent variables looks fairly linear. Thus, our linearity assumption is satisfied.

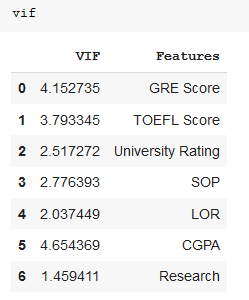
Let us move ahead and check for the multicollinearity.

**Data Scaling**



In this case, we have scaled the data before we check the multicollinearity assumption.





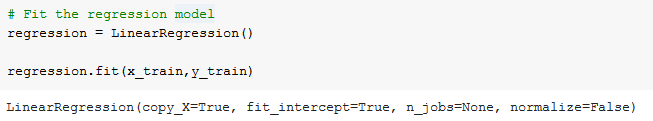
Here, we have added the correlation values for all the features. As a rule of thumb, a VIF value greater than 5 means a very robust multicollinearity. We don’t have any VIF greater than 5, so we are good to go.

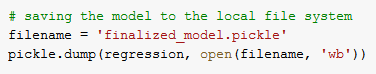
Let us go ahead and use the linear regression and see how good it fits our dataset. But, first, let us split our dataset into train and test.

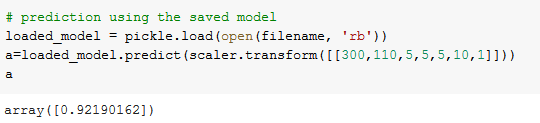


Let’s check the top six observations of y\_train

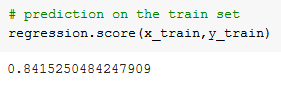




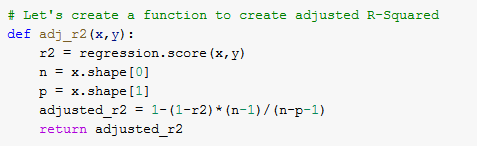




Computing R squared value,



Computing the Adjusted R Squared value,

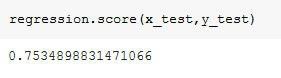




The R Squared score is 84.15% and adjusted R Squared is 83.85% for our training set, which means that we are not being penalized by the use of any feature.

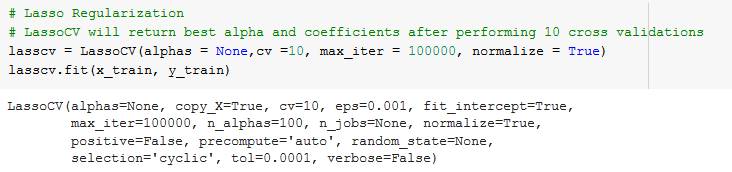
Let’s check how well our model fits the test data.

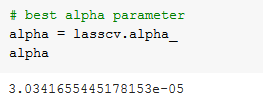
Now let’s check if our model is overfitting the data by using regularization.



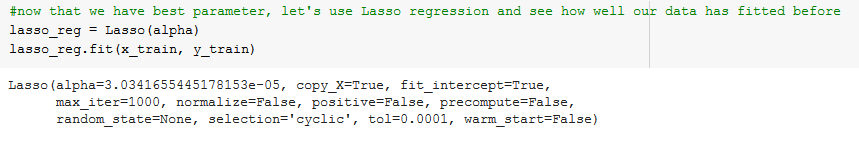
Thus, it looks like our model R squared is less on the test data.

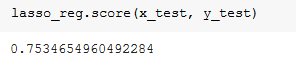
Let’s see if our model is overfitting our training data.





We have derived the best possible alpha value, which is the shrinkage factor for LASSO regression.

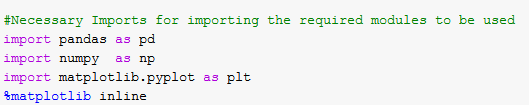




Our R squared for test data (75.34%) comes the same as before using regularization. So, it is fair to say that the OLS model did not overfit the data.

**Task 4) Polynomial Regression**

Importing necessary packages

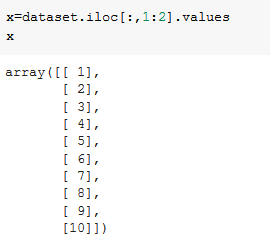


Importing the data





Here, we can see 3 columns in the dataset. The problem statement here is to predict the salary related to the Position and Level of the employee. We may notice that the position and the level are related, as level is an alternate way of conveying the position of the employee in the company. So, essentially the position and level are conveying the same kind of information. As the level is a numeric column, let’s use that in our model. Hence, level is the feature or X variable and salary is a label or the Y variable.





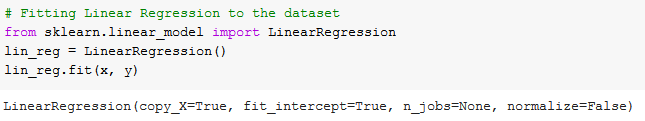


Generally, we divide our dataset into two parts:

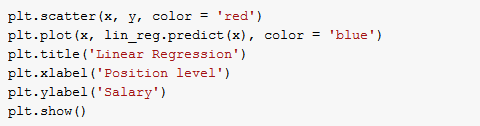
1. The training dataset to train our model.

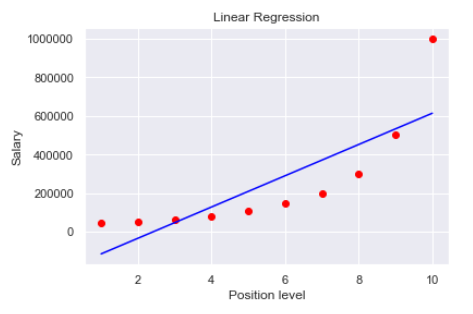
2) The test dataset to test our prepared model.

To learn Polynomial Regression, we take a comparative approach. First, we create a linear model using Linear Regression and then we prepare a Polynomial Regression model and see how the two models compare to each other.

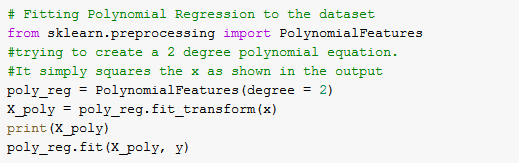


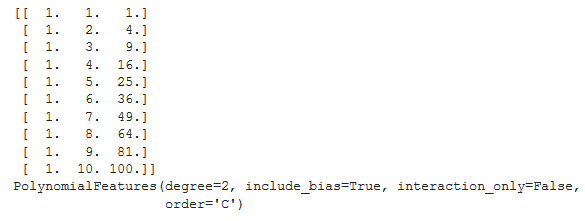
Visualizing the Linear Regression results,

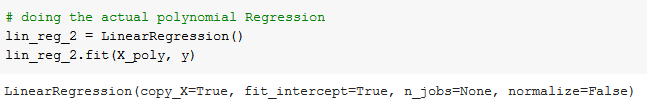




Here, the red dots represent the actual data points, and the blue straight line represents what our model has created. It is evident from the diagram above that a Linear Regression model does not fit our dataset well. Therefore, let us try with a Polynomial model.







As evident, we are using the Linear Regression for the Polynomial Regression as well.

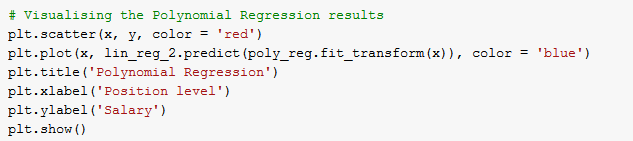
Why is it so?

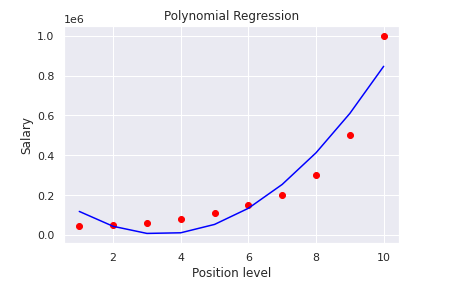
It is because the ‘’Linear’’ in Linear Regression does not consider the degree of the Polynomial equation in terms of the dependent variable(X). Instead, it considers the degree of the coefficients.

Mathematically,

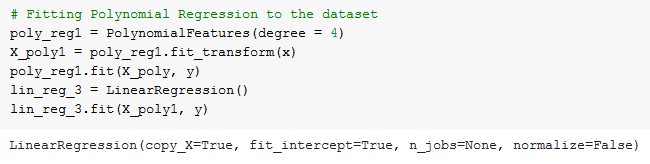


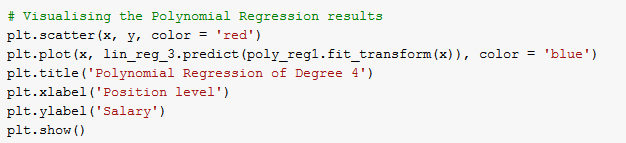
It is considering the power of X, but the powers of a, b, c etc. And as the coefficients are only degree 1, hence the name Linear Regression.

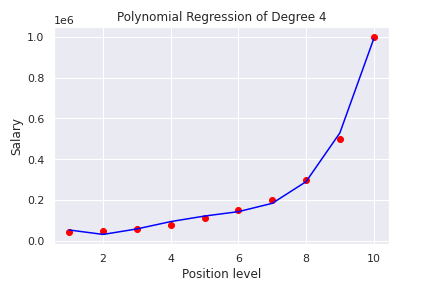




Even so, a second degree is also not a good fit. Now, we will try to increase the degree of the equation *i.e.*, we will try to see whether we get a good fit at a higher degree or not. After some hit and trial, we can observe that the model provides the best fit for the 4th-degree polynomial equation.







Here, we can observe that our model now can accurately fit the dataset. This kind of fit might not be the case with the actual business datasets. We are getting a brilliant fit, partly because of a small number of data points.

## Summary

In this chapter, we have learned the following:

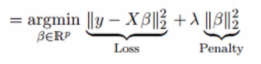
* **Cost function optimization technique,** which helps us to find the parameters with the least/minimum error using the concept of Gradient-Descent technique.
* **Categorical Features in a Regression model**, are handled by converting those categories into the Dummy variables and building the model based on these dummy variables for better computation of the model and avoiding ‘Dummy variable trap’.
* **Feature Scaling** - Two different methods of feature scaling *i.e* Normalization and Standardization. These methods make the data independent of units of measure, which ease the computation and increase the performance of the model.
* **Feature Selection** - We learned about the “*Backward, Forward, and Step-Wise elimination*” techniques to identify the best features and reduce the overfitting problems.
* **Regularization** - Regularization technique is utilized to penalize the parameters responsible for overfitting the model by shrinking the value of the parameters close to or equal to zero, using Ridge, LASSO, and Elasticnet respectively.
* **Polynomial Regression** - Polynomial Regression is employed to capture the non-linear relationship between the dependent and independent variables by including polynomial terms/variables into the model.

## 

## Assessment

## Choose the appropriate option

1. **What is the cost function?**
   1. Difference between predicted and actual value
   2. Difference between the observations
   3. Difference between the variables
   4. None of the above
2. **Optimization technique is used to**
   1. Find the parameters with highest error
   2. Calculate the residuals
   3. Find the parameters with least error
   4. None of the above
3. **Feature scaling is used for**
   1. Regularization of parameters
   2. Making the data independent of units
   3. Calculating the residuals
   4. All of the above
4. **Following is not the feature selection technique**
   1. Forward selection
   2. Backward elimination
   3. ElasticNet
   4. LASSO
5. **From the below formula, lamba is called as**



* 1. Shrinkage parameter
  2. Learning rate
  3. Estimating parameter
  4. None of the above

## Fill in the spaces with appropriate answers

1. Categorical variables are converted to \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.
2. Normalization transforms the values between \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.
3. Gradient Descent is an \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ technique.
4. The Dummy Variables Trap leads to \_\_\_\_\_\_\_\_\_\_\_\_\_\_ in the data.
5. List down the Regularization techniques \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

## True or False

1. Polynomial regression is used to capture the linear relationship between the independent and dependent variables.
   1. True
   2. False
2. Ridge Regression can completely shrink the parameter to Zero
   1. True
   2. False
3. ‘’N’’ number of categories require N-1 Dummy Variables.
   1. True
   2. False
4. Normalization Scales down the values between 0 & 1.
   1. True
   2. False
5. Bias is also called Test Error.
   1. True
   2. False

## Programming Assignment

Using the data in the below URL,

<https://www.kaggle.com/ruiromanini/mtcars>

1. Build a regression model by including the categorical variables
2. Check if the model is *generalized* or not using the regularization techniques.

**Solutions:** Refer to page 47

## Solutions for Assessment

## Choose the appropriate options

1. A
2. C
3. B
4. C
5. A

## Fill in the spaces with appropriate answers

1. Dummy Variables
2. 0 & 1
3. Optimization
4. Multicolinearity
5. Ridge, Lasso, Elasticnet.

## True or False

1. False
2. False
3. True
4. True
5. False

## 

[**CHAPTER 1: INTRODUCTION TO STATISTICS**](#_heading=h.jpx6662nwy0e) **1**

[Theory](#_heading=h.t1r9roulivsb) 1

[Statistics – Big Picture](#_heading=h.ny2aoh92czp3) 1

[Basic Statistical Terminology](#_heading=h.imylnzfub2b6) 1

[Parameter and Statistic](#_heading=h.nv1u40aifso9) 2

[Variables and Data](#_heading=h.t3yhhrpt1p5a) 3

[Variables – Dependent and Independent](#_heading=h.g4xitzfnagy6) 3

[Data](#_heading=h.kshgo37k0mgu) 5

[Levels of Measurement Scales](#_heading=h.o4krhn11b1gh) 5

[Descriptive Statistics](#_heading=h.3143wroeuvwi) 6

[What is a Correlation Matrix?](#_heading=h.dyzsgfg4qek2) 15

[AIM](#_heading=h.5xxv0w1flqqn) 17

[Introduction to Python and Python versions](#_heading=h.y43ezla05mv6) 18

[Anaconda Python Distribution](#_heading=h.jh3s5hxf6jir) 19

[Why Anaconda?](#_heading=h.486740mhpgq4) 19

[Anaconda Installation](#_heading=h.9ak2c3d0exzl) 20

[What are Jupyter notebooks?](#_heading=h.itqk4abc01qw) 24

[How does a notebook work?](#_heading=h.lhth8xt6nabh) 24

[Installing and Launching the Jupyter notebook](#_heading=h.j4a5ab8jwhox) 25

[Launching Jupyter notebook](#_heading=h.aa3w6qqvrgtj) 25

[Importing the necessary packages](#_heading=h.l1ntp6diyqxo) 26

[Calculating Central Tendencies](#_heading=h.ynfc55zgwxvw) 27

[Calculating Measure of Dispersion](#_heading=h.o74uobq3enzm) 27

[Task 2: Standard Deviation](#_heading=h.1hyep9nh992v) 28

[Task 3: Variation](#_heading=h.gvb3ivcgjre4) 29

[Task 4: Range and Percentile](#_heading=h.8vow94og9yv) 29

[Range](#_heading=h.74jkrawy2yn2) 29

[Percentile](#_heading=h.egw8tp2wgh3r) 30

[Task 5: Quartiles](#_heading=h.m8oi9lxb323a) 30

[Task 6: Boxplot](#_heading=h.61kmqrclco7k) 31

[Task 7: Skewness](#_heading=h.2c7qu73ylw9u) 33

[Negatively skewed distribution](#_heading=h.s1g3l7usr4f5) 34

[Positively skewed distribution](#_heading=h.49wskcv38jfg) 35

[Case Study](#_heading=h.y067k1c9sfa9) 35

[Descriptive Statistics](#_heading=h.do0a6dcc2fqf) 38

[Assessment](#_heading=h.dt9w59e8a1j4) 42

[Choose the appropriate option](#_heading=h.xzdlrdgb5ts9) 42

[Fill in the spaces with appropriate answers](#_heading=h.is3edgtrof3k) 44

[Programming Assignment](#_heading=h.f5wlqewl6ybk) 44

[Solutions for Assessment](#_heading=h.64np8u2o7635) 46

[Choose the appropriate options answers](#_heading=h.53skinyo496s) 46

[Fill in the spaces with appropriate answers:](#_heading=h.iq4g32h2qr5j) 46

[Programming Assignment Solution](#_heading=h.xhq19hcypxwv) 46

[**CHAPTER 2: PROBABILITY DISTRIBUTIONS**](#_heading=h.rv03rawki1mu) **48**

[Theory](#_heading=h.1iu570ct3jhx) 48

[Assigning Probabilities](#_heading=h.awjk9xfl5rx2) 48

[Probability Terminology](#_heading=h.9ysk65xlupp9) 51

[Probability – Rules](#_heading=h.w6p8xf5q10ft) 51

[Probability – Types](#_heading=h.bo60im4j1d66) 53

[Probability Distributions](#_heading=h.vxy842ge041) 57

[Types of Discrete Probability Distributions](#_heading=h.wzuab14khsrp) 60

[Poisson distribution Vs Binomial distribution](#_heading=h.la3mdos3yu6g) 65

[AIM](#_heading=h.kc791q66cbez) 66

[Importing the necessary packages](#_heading=h.2jctls4i2ubm) 68

[Calculating probability using binomial distribution](#_heading=h.324z8iren7td) 68

[Calculating the probability using Poisson distribution](#_heading=h.p5ik1o72u4i8) 73

[Case Study](#_heading=h.p9ec32t7m77g) 75

[Assessment](#_heading=h.3ipto37teh05) 77

[Choose the appropriate option](#_heading=h.9uodfqflk2jl) 77

[Fill in the spaces with appropriate answers](#_heading=h.avr1bf3vgc8p) 78

[Programming Assignment](#_heading=h.9lrfjjq2u6uu) 78

[Solutions for Assessment](#_heading=h.wgspyzk9001a) 80

[Choose the appropriate options answers](#_heading=h.sqm45mm41pbe) 80

[Fill in the spaces with appropriate answers](#_heading=h.ntqxws1x0rnp) 80

[Programming Assignment Solutions](#_heading=h.2tgcx6rr26k3) 80

[**CHAPTER 3 – INFERENTIAL STATISTICS**](#_heading=h.s31uf5pf2iws) **81**

[Theory](#_heading=h.4v9i2459g7q6) 81

[Continuous Random Variable](#_heading=h.ljnbbxu8lvwx) 81

[Types of Continuous Random Variables](#_heading=h.o1dbxusubnr8) 82

[Sampling Distribution](#_heading=h.3bl7goup0vgb) 93

[Sampling Distribution of the Means](#_heading=h.sy6qwajpxb4e) 93

[Expectation and Standard deviation of 𝑋 bar](#_heading=h.59cgh5j3n0md) 94

[Two Main areas of inferential statistics](#_heading=h.spn7f8gbm2cj) 94

[Confidence Intervals of Mean, Proportion, and Variance](#_heading=h.j6sahogb1vw7) 97

[Mean:](#_heading=h.yd3onbxpnybi) 97

[Proportion:](#_heading=h.b99cd2sgrhn9) 99

[t-Distribution:](#_heading=h.lq6xl63qdnc2) 100

[Chi Square Distribution for Variance:](#_heading=h.h9v8a3c36i1s) 102

[Confidence Intervals Formulas](#_heading=h.mq1htd8cu5b0) 105

[Critical Region up Close](#_heading=h.grafeotblxor) 109

[One-tailed tests:](#_heading=h.lfljjlu1c8j4) 109

[Two-tailed tests:](#_heading=h.pl525og96iud) 110

[Types of Errors](#_heading=h.x9nnfm87udt1) 110

[Type I error:](#_heading=h.pnaq195vnaba) 111

[Type II error:](#_heading=h.54b9e6pqmgk0) 112

[Why are Type I and Type II Errors Important?](#_heading=h.rlbb1jwxunlc) 112

[Probabilities of Errors in our Example:](#_heading=h.l3t4zto27ucm) 113

[Power of Hypothesis Test:](#_heading=h.ev8y6u8zw2eb) 114

[Common Test Statistics for Inferential Techniques](#_heading=h.or2xmsudcky3) 115

[Types of t – tests](#_heading=h.4flv3trjt0ct) 115

[Two-Sample t-test for Paired Data:](#_heading=h.wp1p8uy5c5m4) 115

[Two-Sample t-Test for Unpaired Data:](#_heading=h.pjj6l2a1h69s) 118

[Confidence Intervals](#_heading=h.x2tlldvqw4hv) 122

[AIM](#_heading=h.b61pw5u5t0f2) 123

[Importing the necessary packages](#_heading=h.xpus6lc8pxf0) 125

[Calculating probabilities using normal and standard normal distributions](#_heading=h.m4sruomotkbx) 125

[Calculating the confidence intervals](#_heading=h.ltnlh147piu2) 134

[Performing Test of Hypothesis](#_heading=h.qtwui8mbr5md) 139

[Assessment](#_heading=h.4ov04yvcjmu7) 147

[Choose the appropriate option](#_heading=h.l2qb4mlwjct2) 147

[Fill in the spaces with appropriate answers](#_heading=h.vhgminvoejwe) 148

[Programming Assignment](#_heading=h.1mmjp4z0spj8) 148

[Solutions for Assessment](#_heading=h.evtxl9by0v3r) 149

[Choose the appropriate options](#_heading=h.oym0obdf444y) 149

[Fill in the spaces with appropriate answers](#_heading=h.rzvc0miialvz) 149

[Programming Assignment Solution](#_heading=h.k68dbndxtj92) 149

[**CHAPTER 4: ANOVA**](#_heading=h.vwmktrwlb2qd) **150**

[**Analysis of Variance**](#_heading=h.x0w8ry9edtbn) **150**

[Theory](#_heading=h.u57ak69yg9ee) 150

[Assumptions for ANOVA](#_heading=h.rfib7jgos4pw) 150

[Why ANOVA?](#_heading=h.7vv45j3ja4mv) 151

[Hypothesis test for 2 sample variances](#_heading=h.xt4n7tmm04b) 152

[Data](#_heading=h.tqrleqc9415f) 153

[What is a null and alternate hypothesis?](#_heading=h.p7e2ngq2oel9) 153

[What are numerator and denominator degrees of freedom?](#_heading=h.rrgx47xvxmq4) 153

[Reading an F-table](#_heading=h.qd53x7k0xzu0) 154

[Will you reject the null hypothesis or not?](#_heading=h.eqwcayox4d1m) 155

[What are the business implications?](#_heading=h.crw4fxln9v9) 155

[Comparing means](#_heading=h.7ahw0fyaq3e1) 155

[Comparing with population mean](#_heading=h.6ui3agpp687x) 155

[Why can’t we compute the t-test three times?](#_heading=h.wt1rnns477za) 156

[What changed?](#_heading=h.wz7igoj2gfkp) 157

[ANOVA: Analysis of Variance is variability ratio.](#_heading=h.n576n93o2k35) 158

[Solutions:](#_heading=h.c2sed0l3ps3u) 172

[Null and Alternative hypothesis](#_heading=h.tj6dh669cri1) 172

[Observed Frequencies](#_heading=h.tj9i3s25nbwe) 173

[Expected Frequencies](#_heading=h.69cm0xb2oao4) 173

[Chi-square](#_heading=h.laz4tapmduui) 173

[Case Study](#_heading=h.3vdjpqewznr0) 175

[Task 2: Two way ANOVA](#_heading=h.8lak4wcbuuo5) 182

[Task 3: Chi-square test of independence](#_heading=h.ctlu8qi4zu6k) 183

[Assessment](#_heading=h.4qlcgxln07b7) 185

[Choose the most appropriate option](#_heading=h.3os0llhyzm15) 185

[Fill in the spaces with appropriate answers](#_heading=h.hnvd5zmjvbtv) 186

[Programming Assignment](#_heading=h.ypaoe327k3n2) 186

[Solutions for Assessment](#_heading=h.cv2024jcexba) 187

[Choose the appropriate options](#_heading=h.p7r3npc0pdkg) 187

[Fill in the spaces with appropriate answers](#_heading=h.j9dl4lm3o7cp) 187

[Programming Assignment Solution](#_heading=h.d59djkifgpif) 187

[**CHAPTER 5: Data Cleaning and Exploratory Data Analysis**](#_heading=h.el54ufhbqwm0) **188**

[Theory](#_heading=h.ngvz7inggp82) 188

[Case Study](#_heading=h.aspoem61k4d4) 188

[Problem Definition](#_heading=h.15burvkdeb3) 188

[Imputing Missing Values](#_heading=h.ogk69wkz2oo4) 194

[Univariate analysis (Single variable plots)](#_heading=h.671mvg3wck24) 196

[Removing Outliers](#_heading=h.5qt4tnj08gb) 200

[Bivariate Analysis](#_heading=h.ch6yxg6ci5v5) 201

[Correlations between Features and Target](#_heading=h.i9zco8xp4bve) 204

[Multivariate Analysis](#_heading=h.otxbdi9yaka0) 208

[Assessment](#_heading=h.hffziwfwdj77) 215

[Choose the appropriate option](#_heading=h.2p113gwgrpc9) 215

[Programming Assignment](#_heading=h.t3gbyk26ujei) 216

[Solutions for Assessment](#_heading=h.957nifucmkve) 217

[Choose the appropriate options](#_heading=h.epehh2a4egah) 217

[Programming Assignment Solution](#_heading=h.wyln5r7ax4ml) 217

[**CHAPTER 6: Regression Analysis - Part I**](#_heading=h.90agk616jlca) **220**

[Theory](#_heading=h.3ngr6gldjbmv) 220

[The use of Regression Analysis](#_heading=h.k2y0axdw43rd) 220

[Linear Regression](#_heading=h.dwkx3kcdwbbo) 221

[Problem Statement](#_heading=h.g2sjxhnqyr16) 222

[Ordinary Least Squared Estimation](#_heading=h.ywd38pg4tt1a) 225

[The Mathematics involved](#_heading=h.xdms2js53ch5) 225

[Derivation of OLS by Minimizing Errors](#_heading=h.zddutj4hwv7p) 226

[Interpreting the model](#_heading=h.urq8g560r5qk) 227

[Prediction using the model](#_heading=h.1f2cj1b5ea67) 227

[Model Confidence](#_heading=h.2hvsdv52kbi7) 229

[Hypothesis Testing and p-values](#_heading=h.g0c6glicu8mh) 230

[Assumptions of OLS Regression](#_heading=h.7o2reruim5og) 233

[What if these assumptions are violated?](#_heading=h.d82hvdt51xp4) 234

[Interpretation of Regression Plots](#_heading=h.9t1ri98nlit9) 235

[Residual vs Fitted Values](#_heading=h.6m5as85qznxr) 236

[Normal Q-Q Plot](#_heading=h.e9ptkf64knw) 238

[Scale Location Plot](#_heading=h.goqzkga7snom) 238

[Residuals vs Leverage Plot](#_heading=h.tu9ieh1i8h8i) 239

[Model Evaluation Metrics](#_heading=h.23qlp9t9fk03) 240

[R Square/Adjusted R Square](#_heading=h.gnk1gx4v191u) 240

[Adjusted R Square](#_heading=h.px0s9migmuxv) 241

[Mean Square Error (MSE)](#_heading=h.wx8ay04giwgz) 242

[Root Mean Square Error (RMSE)](#_heading=h.jt6art4awg24) 243

[Mean Absolute Error](#_heading=h.wubkcdkni9lr) 243

[Overall Recommendation/Conclusion](#_heading=h.o3ia9qq5lkng) 244

[Practical Implementation](#_heading=h.aq2v33a7hiez) 244

[Import our libraries](#_heading=h.ar9l2n58v3k7) 244

[Load the Data into Pandas](#_heading=h.dgc0bcu3n2ti) 244

[Check for Multicollinearity](#_heading=h.isd6xxlsj6sh) 247

[Describe the Dataset](#_heading=h.4xp1lxa2e509) 250

[Filtering the Dataset](#_heading=h.1b8k6xcusltf) 251

[Build the model](#_heading=h.ag5x30oovu76) 251

[Exploring the output](#_heading=h.2hy01dvk0qcl) 252

[Evaluating the model](#_heading=h.hi0s12rjdsua) 253

[Using the Statsmodel](#_heading=h.z4n4rok5gaux) 253

[Checking for HeteroScedasticity](#_heading=h.rjcn73y7z1h) 254

[Checking for Autocorrelation](#_heading=h.cu4g8gcd6stq) 255

[Checking for Normally Distributed Residuals](#_heading=h.kvvx8cdt09gn) 257

[Checking the Mean of the Residuals Equals 0](#_heading=h.dnhlbzogmzhz) 257

[Measures of Error](#_heading=h.4h5rimd178dv) 258

[R-Squared](#_heading=h.ers182hngapj) 259

[Confidence Intervals](#_heading=h.gi2c348ghjq2) 259

[Hypothesis Testing](#_heading=h.sdp3ckvzq51) 260

[Create a Summary of the Model Output](#_heading=h.sh5adwty45mt) 261

[Remove the Insignificant Variables](#_heading=h.m7najmks77xx) 263

[Save the Model for Future Use](#_heading=h.1dde7vw6u210) 264

[Summary](#_heading=h.dgtwun7tirjl) 265

[Assessment](#_heading=h.45f7c8iu1r6b) 266

[Choose the appropriate option](#_heading=h.af9f58yeweal) 266

[Fill in the spaces with appropriate answers](#_heading=h.8f3y2mh0d74v) 267

[True or False](#_heading=h.37t8lnwq80m) 267

[Programming Assignment](#_heading=h.ca6ynyw10ecj) 267

[Solutions for Assessment](#_heading=h.ft0uvcw14ghu) 268

[Choose the appropriate options](#_heading=h.26ulotynbt8b) 268

[Fill in the spaces with appropriate answers](#_heading=h.l8vsr5m8ynj6) 268

[True or False](#_heading=h.xgy03k12ejv4) 268

[**CHAPTER 7: Regression Analysis – Part II**](#_heading=h.65yziot7xayx) **269**

[Theory](#_heading=h.o2p3ggpk83fr) 269

[Cost function Optimization Algorithm](#_heading=h.qe52eqe6yen4) 269

[What is Gradient Descent?](#_heading=h.quje536mafjh) 269

[Dummy variables versus quantitative explanatory variable](#_heading=h.xuq8z53xzjt1) 280

[Why do we need to perform feature scaling?](#_heading=h.wg00m685putq) 281

[Feature scaling techniques](#_heading=h.hdvy9frpqolb) 281

[Normalization](#_heading=h.da3d8m896rz7) 281

[Standardization](#_heading=h.625awf33w9bf) 282

[Backward Elimination](#_heading=h.uzkqrofagaxr) 283

[Forward Selection](#_heading=h.y9j38ewnczq) 284

[Stepwise Regression](#_heading=h.xaka3qdx523r) 284

[Underfitting and Overfitting](#_heading=h.3ngbcd1d51d9) 285

[Underfitting](#_heading=h.frdvxu791buo) 285

[Overfitting](#_heading=h.kyfzf7oha9wi) 285

[Bias and Variance Tradeoff](#_heading=h.md6ls9k4dxhz) 286

[Regularization](#_heading=h.2wmjyhffhnh2) 288

[Polynomial Regression](#_heading=h.2v7doovvgadn) 291

[When to use Polynomial Regression?](#_heading=h.ab0gdmp5kyty) 292

[AIM](#_heading=h.9px0slnv6wu7) 294

[Task 1): Handling Categorical Predictors with Two Categories](#_heading=h.8hf4d19xnda) 295

[Task 2) Handling Categorical Variables with More Than Two Categories](#_heading=h.3s4qd410g05a) 297

[Task 3) Regularization the algorithms](#_heading=h.ik4w6nnxjgrk) 299

[Summary](#_heading=h.o3s3pxra2vfv) 313

[Assessment](#_heading=h.j03a5h9oql2t) 314

[Choose the appropriate option](#_heading=h.id74mxobg9o4) 314

[Fill in the spaces with appropriate answers](#_heading=h.hdw1407ws2n4) 315

[True or False](#_heading=h.xtbw63qr13ws) 315

[Programming Assignment](#_heading=h.o8g65zicqpa1) 315

[Solutions for Assessment](#_heading=h.vlxdtyx0oyy4) 316

[Choose the appropriate options](#_heading=h.ouajadfgto01) 316

[Fill in the spaces with appropriate answers](#_heading=h.lpmdrx9p372j) 316

[True or False](#_heading=h.9c1vkjlxxscw) 316

## 

* <https://en.wikipedia.org/wiki/Statistics>
* <https://docs.anaconda.com/anaconda/install/windows/>
* <https://docs.python.org/3/>
* <https://numpy.org/doc/stable/reference/>
* <https://docs.scipy.org/doc/scipy/reference/stats.html>
* <https://en.wikipedia.org/wiki/Probability_distribution>
* <https://en.wikipedia.org/wiki/Normal_distribution>
* <https://en.wikipedia.org/wiki/Normal_distribution#Standard_normal_distribution>
* <https://en.wikipedia.org/wiki/Confidence_interval>
* <https://en.wikipedia.org/wiki/Statistical_hypothesis_testing>
* <https://en.wikipedia.org/wiki/Analysis_of_variance>
* <https://en.wikipedia.org/wiki/Chi-squared_test>
* [https://matplotlib.org/stable/index.html#](https://matplotlib.org/stable/index.html)
* <https://seaborn.pydata.org/>
* <https://scikit-learn.org/stable/modules/generated/sklearn.impute.SimpleImputer.html>