NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

SC1015 : Review Lecture

Basic Statistics
and some EDA

Dr Smitha K G

SC1015

Admin Announcements

1. Detailed solutions to the LAMS Quizzes will be posted every week. Please try on your own first.
2. Detailed solutions to the Lab Exercises will be posted every week after the Lab Week is over.

LAMS Completion Status

Module 1 Part 1 : Above 900 – Quiz solutions posted

Module 1 Part 2 : Around 800 – Quiz solutions posted

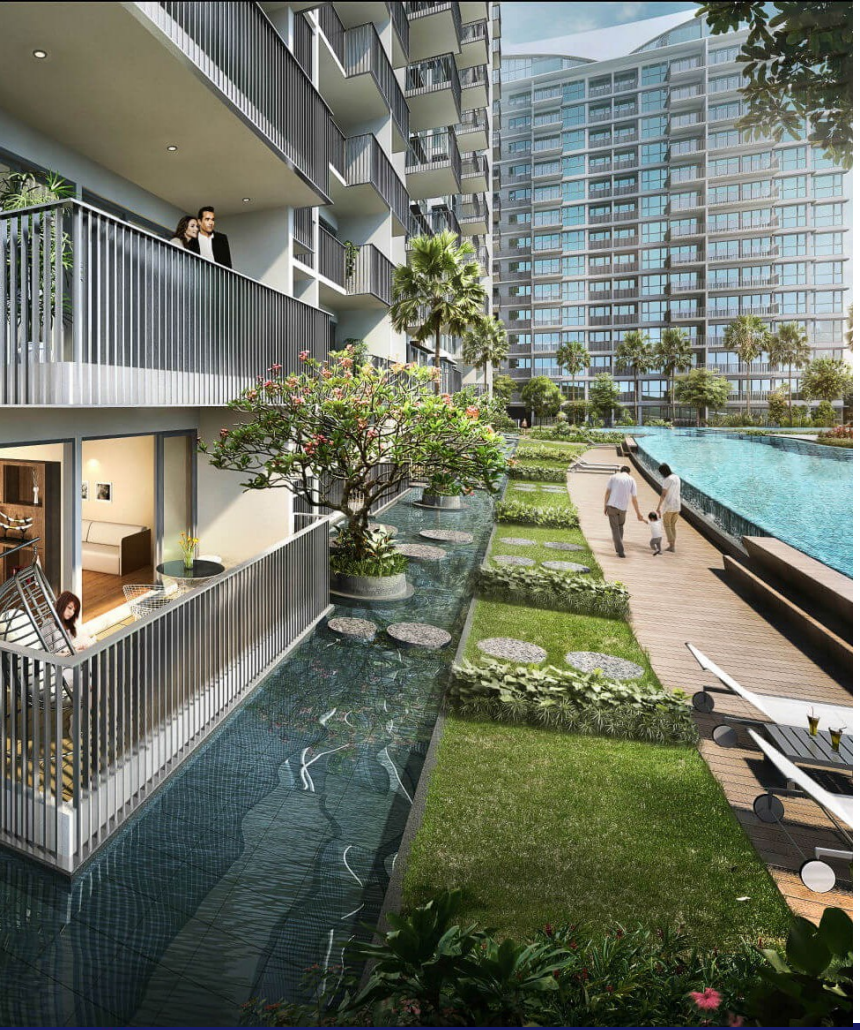
Module 2 Part 1 : Above 500 – needed for Ex2

Module 2 Part 2 : Around 260– Complete by Exercise 3 in W4

Lab 2- Basic Statistics on Week 3

Graded Lab Exercise on EDA – in Week 4.

Please note that Friday 9th FEB- no lab classes- Lab 3 for them will be on 16th FEB



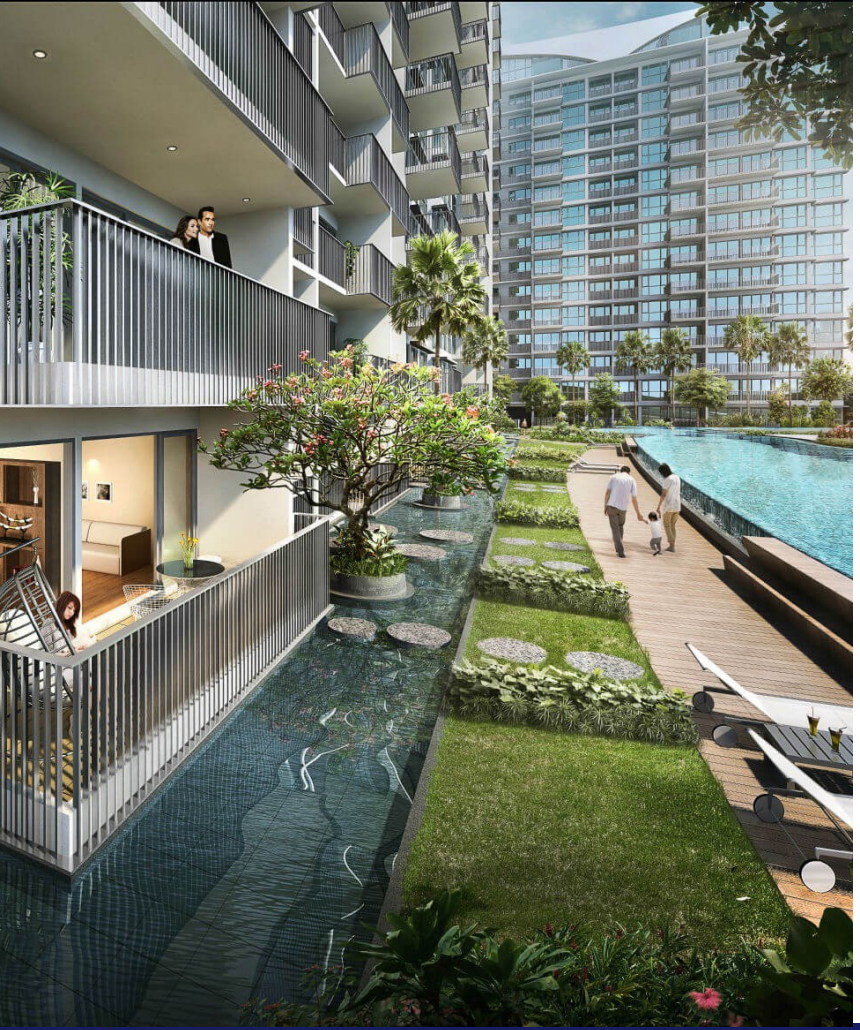
Data Science

Problem Formulation

Suppose that one of your close relatives is planning to buy a Condo in Singapore.

Devise a strategy for them to judge if the price quoted by Seller is expected, higher than expected, or lower than expected, for any given Condo.

How would you perform the Sample Collection?
Is there any requirement for Data Preparation?
How would you finally Formulate the Problem?



Data Science

Estimation Strategy

Which one of the methods do you prefer?

1. **Compare prices of “similar” condos to estimate the price of your target condo.**
2. **Build a “formula” to estimate price of a condo given its features and attributes.**

There are several Model Families in data science.

Choice 1 will imply **Nearest Neighbors**.

Choice 2 will imply **Regression Models**.

https://scikit-learn.org/stable/supervised_learning.html

Getting Started Prediction Competition

House Prices - Advanced Regression Techniques

Predict sales prices and practice feature engineering, RFs, and gradient boosting

Kaggle · 6,330 teams · Ongoing

Overview | Data | Notebooks | Discussion | Leaderboard | Datasets | Rules | Join Competition

Overview

- Description
- Evaluation
- Tutorials
- Frequently Asked Questions

Start here if...

You have some experience with R or Python and machine learning basics. This is a perfect competition for data science students who have completed an online course in machine learning and are looking to expand their skill set before trying a featured competition.

Competition Description

Ask a home buyer to describe their dream house, and they probably won't begin with the height of the basement ceiling or the proximity to an east-west railroad. But this playground competition's dataset proves that much more influences price negotiations than the number of bedrooms or a white-picket fence.

With 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, this competition challenges you to predict the final price of each home.

Practice Skills

- Creative feature engineering
- Advanced regression techniques like random forest and gradient boosting

Acknowledgments

The Ames Housing dataset was compiled by Dean De Cock for use in data science education. It's an incredible alternative for data scientists looking for a modernized and expanded version of the often cited Boston Housing dataset.

Notebooks >

873 discussion topics >

Data Science

Machine Learning

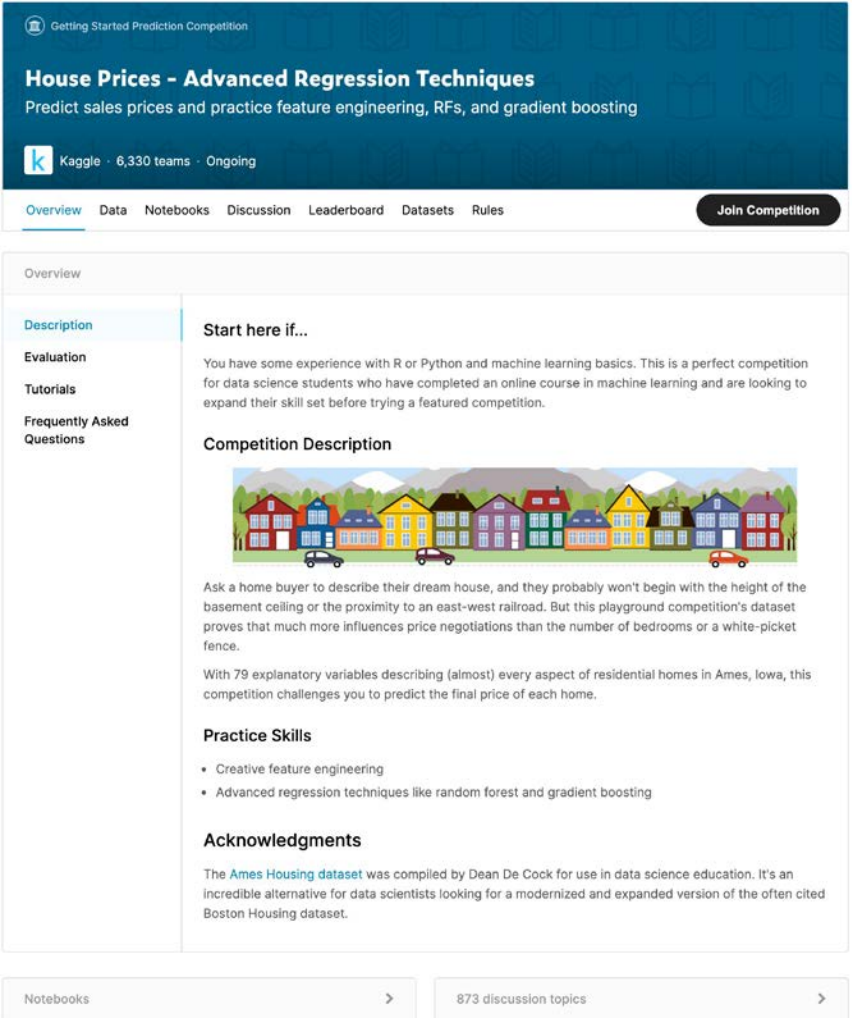
Prediction : Numeric

Regression

Model : SalePrice = *f* (Variables)

Given	Some Houses as Train Data
Learn	The Formula for SalePrice
Predict	Estimate SalePrice for Test

https://www.kaggle.com/c/house-prices-advanced-regression-techniques



Data Science

Exploratory Data Analysis

Target : Predicting SalePrice

Model : SalePrice = *f* (Variables)

Understand the variable **SalePrice**
Understand **all** the other Variables
Understand all **mutual** relationships
Clean and prepare the Training Data
Then think of creating your **Model** ;-)

https://www.kaggle.com/pmarcelino/comprehensive-data-exploration-with-python

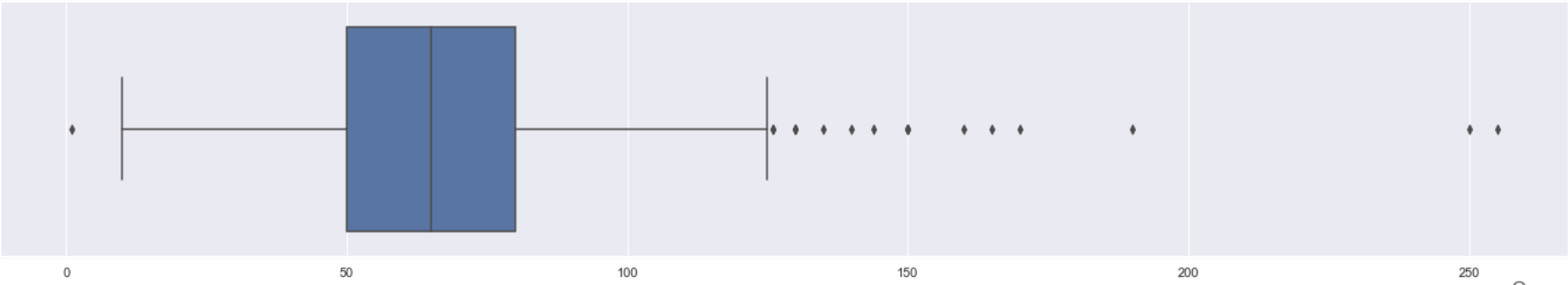
Let’s clarify a few points on …

BASIC STATISTICS

Uni-Variate Box-Plot

min	1.000000
25%	50.000000
50%	65.000000
75%	80.000000
max	255.000000

Helps us to get a glimpse of the data distribution and spot the outliers.



1. Question 7

Suppose the median household income in Singapore is SGD 9,000, and the quartiles are Q1 = SGD 4,500 and Q3 = SGD 10,500. What can you infer about the outliers? *

Note : Multiple answers may be correct. Select all the ones you think are right.

Choose at least one answer.

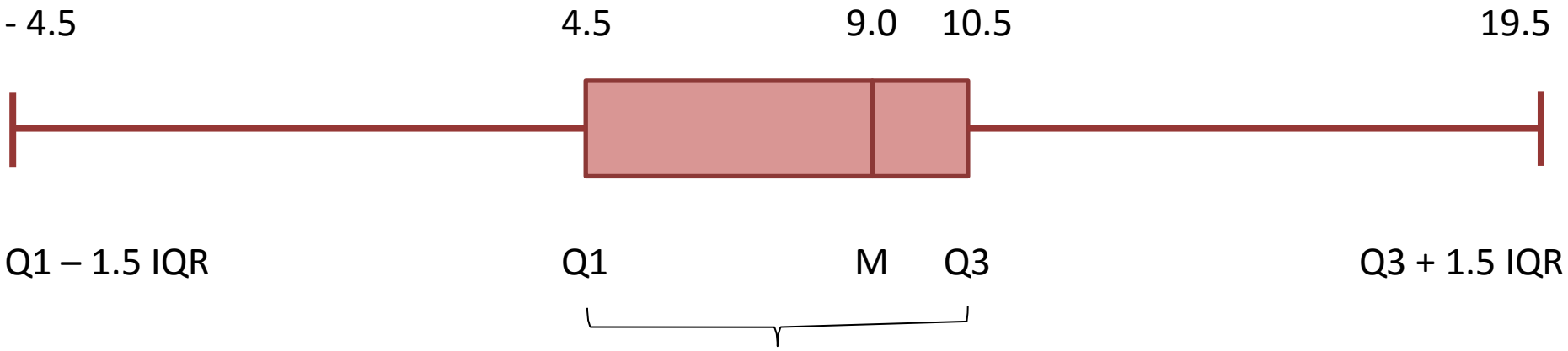
- [ ] Household income less than SGD 1,000 may be considered as outliers (abnormally low) in this data.
- [x] Household income above SGD 20,000 may be considered as outliers (abnormally high) in this data.

  That's quite far indeed. It's higher than the third quartile by around 6 quartile gaps on the higher side. That may be considered outlier.
- [ ] Household income above SGD 10,500 may be considered as outliers (abnormally high) in this data.
- [ ] We can't say that any of the other answers are true unless we know the average household income.

Marks for this submission: 1/1.

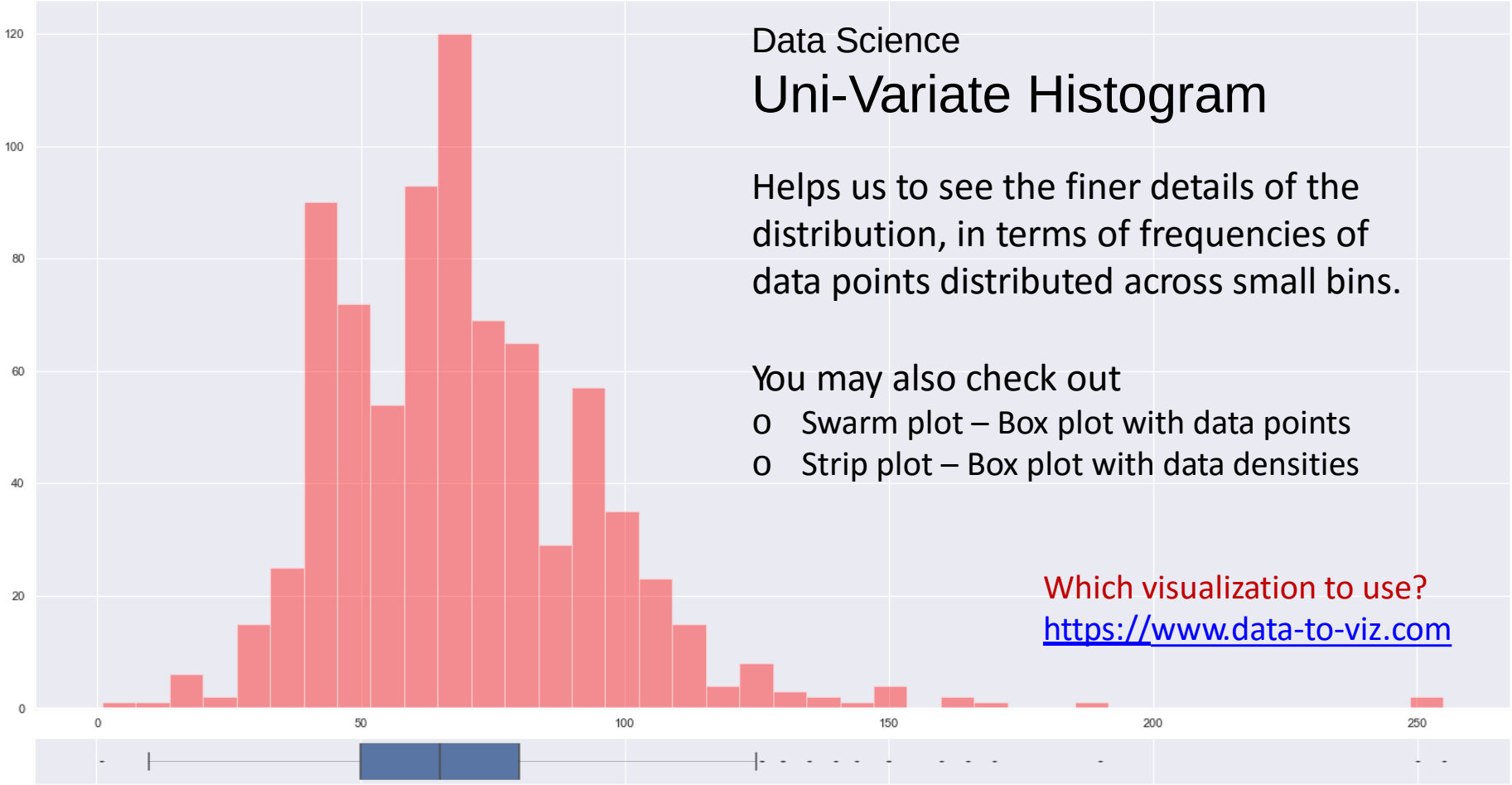
M2 Part 1
LAMS Question

Median household income in Singapore is **SGD 9,000**
Quartiles are Q1 = **SGD 4,500** and Q3 = **SGD 10,500**
What can you infer about the Outliers in this case?



Inter-Quartile Range (IQR) = 10,500 – 4,500 = 6,000

What do you do with the outliers in data? – We will consider this again in Regression
How about missing values in the data? – Do you drop them, fill them, or predict them?



Data Science

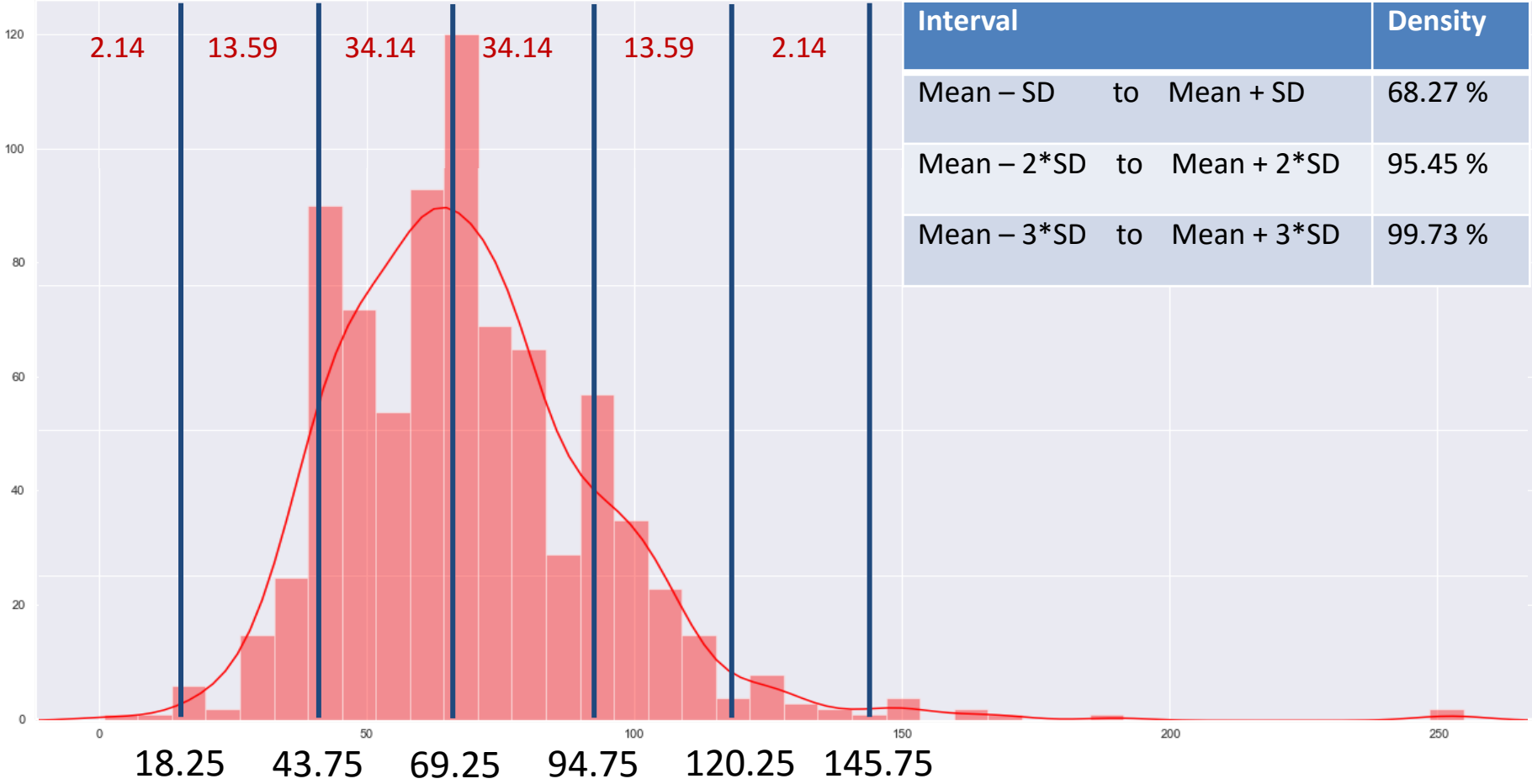
Uni-Variate Histogram

Helps us to see the finer details of the distribution, in terms of frequencies of data points distributed across small bins.

You may also check out

- Swarm plot – Box plot with data points
- Strip plot – Box plot with data densities

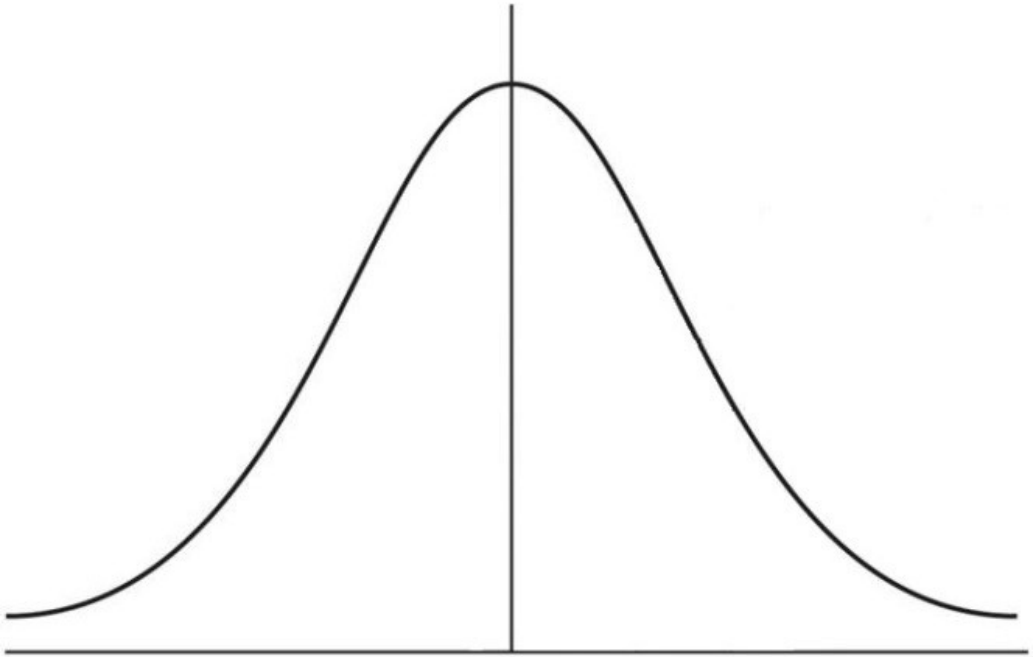
Which visualization to use?
https://www.data-to-viz.com



Interval			Density
Mean – SD	to	Mean + SD	68.27 %
Mean – 2*SD	to	Mean + 2*SD	95.45 %
Mean – 3*SD	to	Mean + 3*SD	99.73 %

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

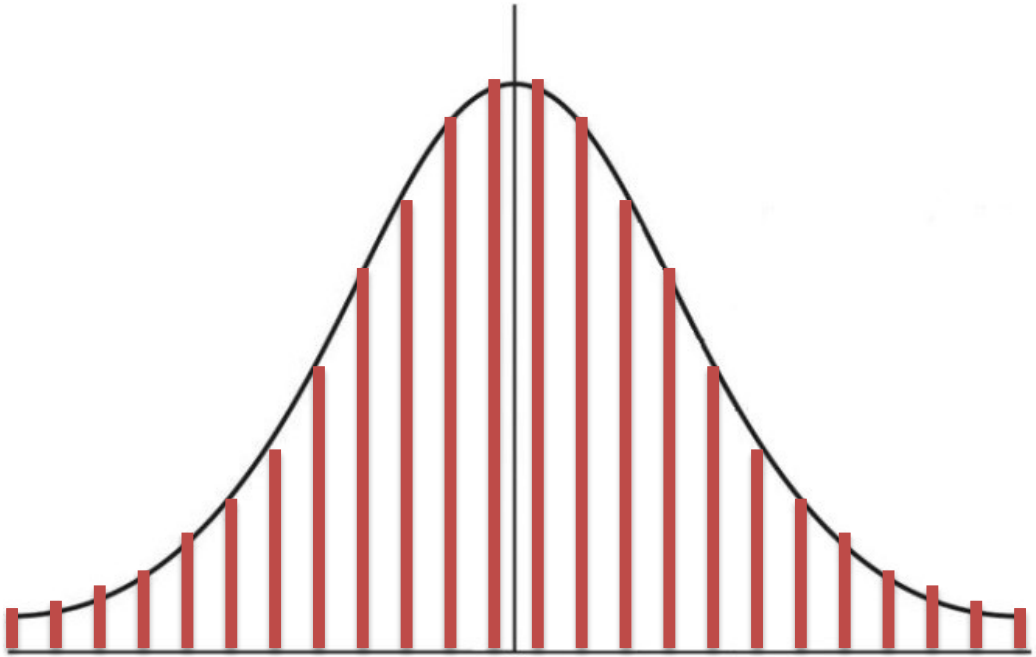
Probability Density Function (PDF)



https://en.wikipedia.org/wiki/Normal_distribution

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Probability Density Function (PDF)

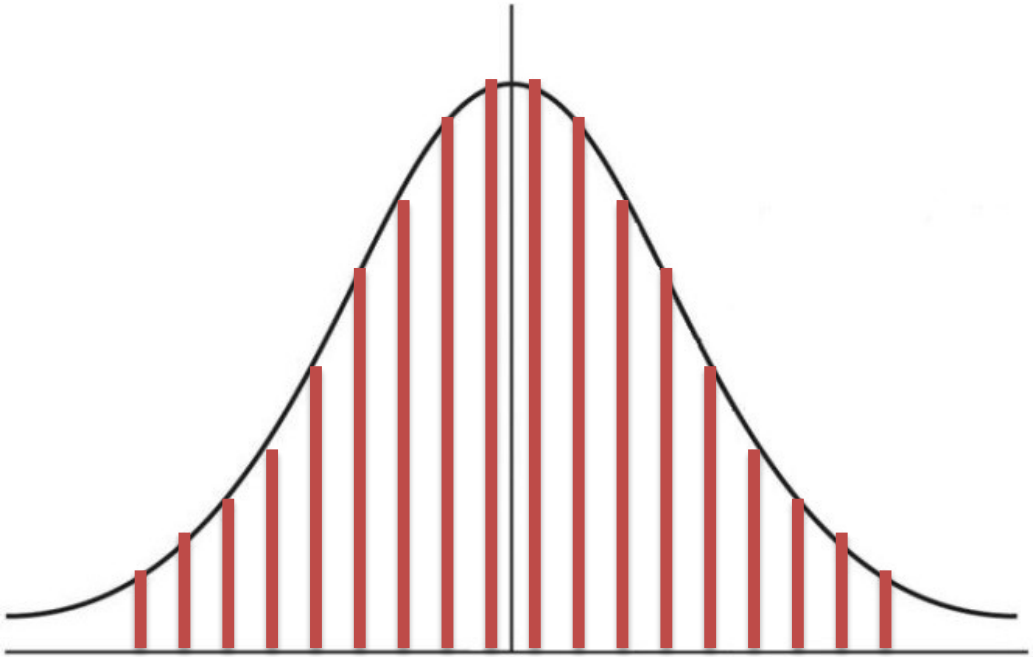


$$\int_{-\infty}^{\infty} f(x)\, dx = 1$$

https://en.wikipedia.org/wiki/Normal_distribution

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Probability Density Function (PDF)

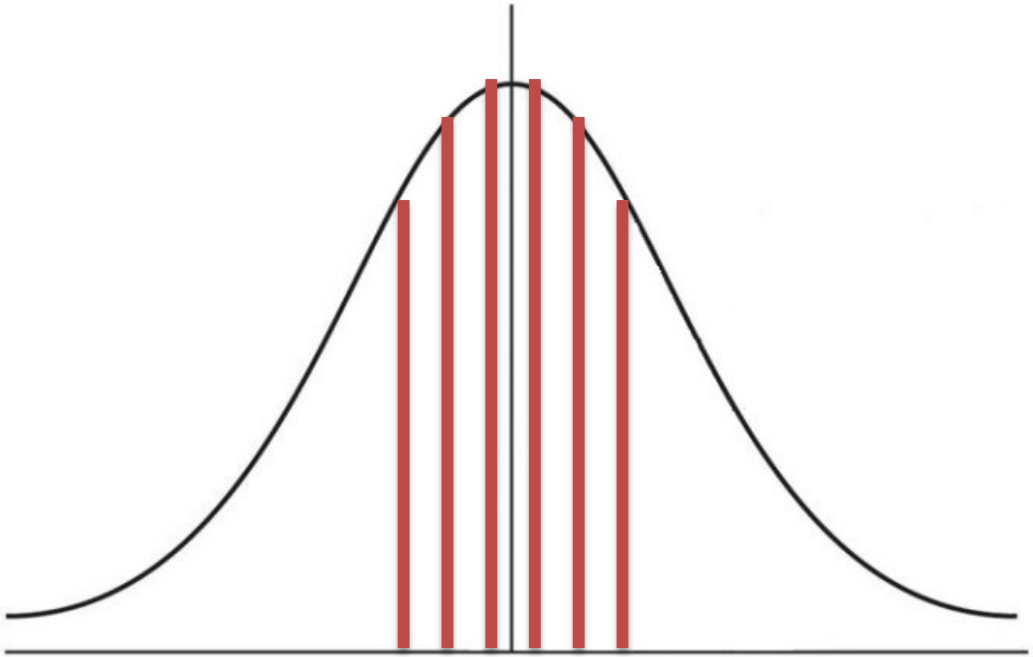


$$\int_{\mu-3\sigma}^{\mu+3\sigma} f(x)\,dx \approx .997$$

https://en.wikipedia.org/wiki/Normal_distribution

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Probability Density Function (PDF)

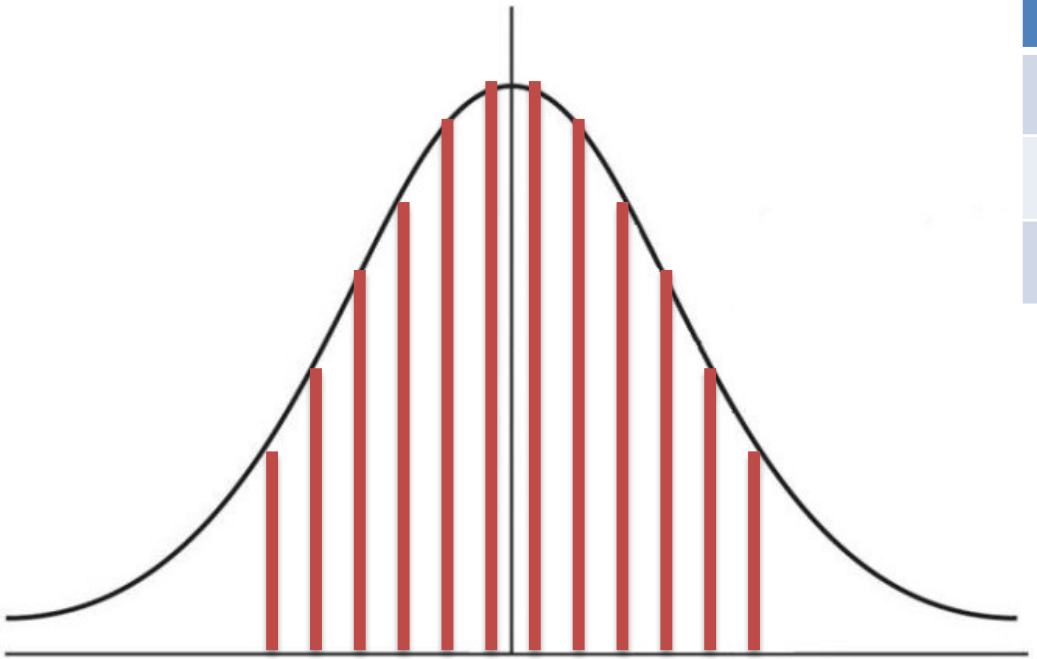


$$\int_{\mu-\sigma}^{\mu+\sigma} f(x)\, dx \approx .683$$

https://en.wikipedia.org/wiki/Normal_distribution

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

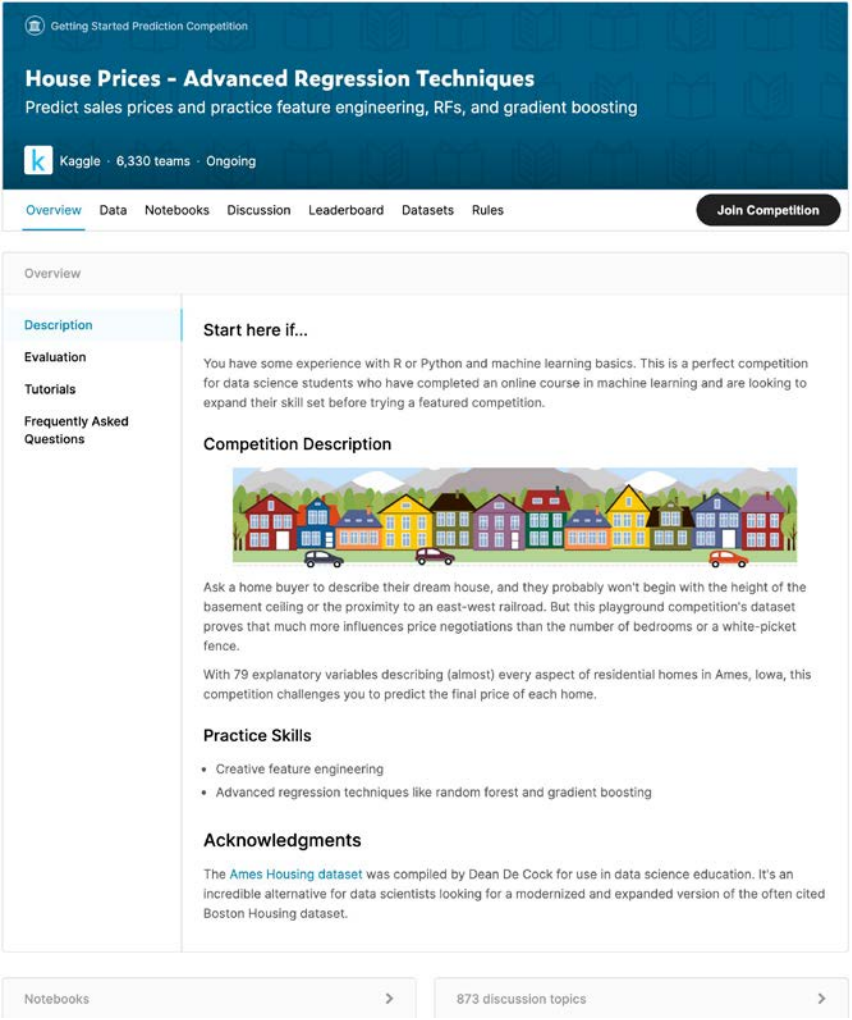
Probability Density Function (PDF)



Confidence	Interval in terms of SD
90 %	Mean +/- 1.645 * SD
95 %	Mean +/- 1.96 * SD
99 %	Mean +/- 2.576 * SD

Well, I am 95% confident that the sale price of this condo should be within SGD 650K and SGD 925K.

https://en.wikipedia.org/wiki/Normal_distribution



Data Science

Do I care about Normality?

Target : Predicting SalePrice

Model : SalePrice = *f* (Variables)

You generally **estimate** the Mean

Estimation works **better** for Normal

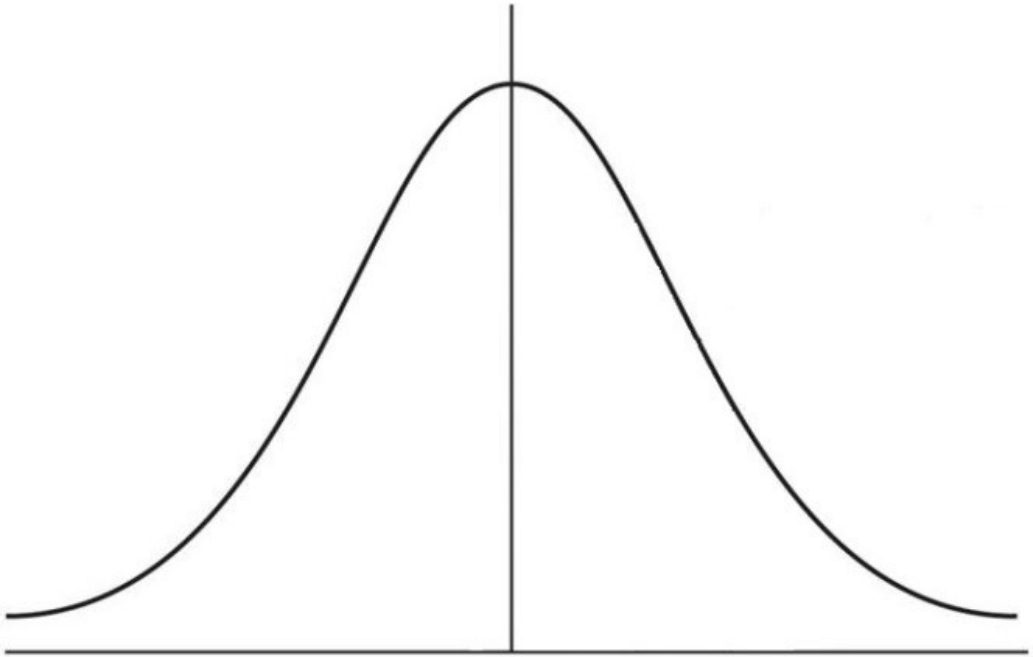
Check if SalePrice **looks like** Normal

If not, is it possible to **Transform** it?

https://developers.google.com/machine-learning/data-prep

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Probability Density Function (PDF)



Described by two parameters: Mean and Standard Deviation

- Can you describe a distribution just by one parameter, Mean?
- What if you can't describe a distribution by Mean and SD?

https://en.wikipedia.org/wiki/Normal_distribution

Moments of a Function

Moments are quantitative measures related to the shape of a distribution.

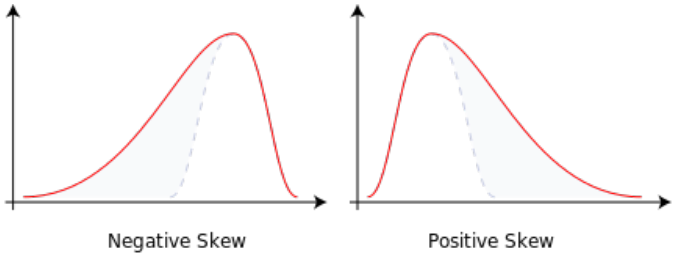
$(x-\mu)^2$ for the *n*-th moment

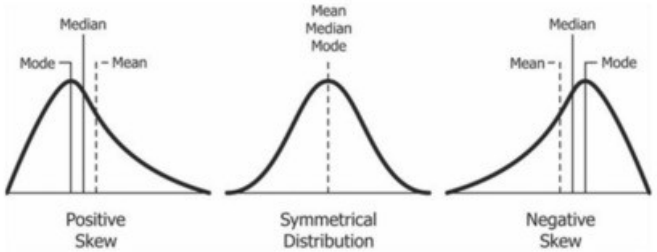
First Moment	: Mean	(centrality)
Second Moment	: Variance	(dispersion)
Third Moment	: Skewness	(asymmetry)
Fourth Moment	: Kurtosis	(heaviness)

https://en.wikipedia.org/wiki/Moment_(mathematics)

Asymmetric Distributions

Skewness and Kurtosis





We want some variables to be symmetric.

Third Moment : Skewness (asymmetry)

Fourth Moment : Kurtosis (heaviness)

In case the distribution is too skewed, you may want to make it more normal by applying standard transformations.

Transforms : Logarithm, Exponential, Square Root, Square

https://www.kaggle.com/getting-started/110134
https://towardsdatascience.com/top-3-methods-for-handling-skewed-data-1334e0debf45

Xtra Module : Non-Examinable LAMS Sequences

The LAMS Sequences in this folder are NOT examinable, and has no quiz. So, there is NO deadline.
Feel free to go through them (only if you are interested) to learn more about additional Data Science topics.

LAMS Quizzes : Questions, Answers and Explanations

Attached Files: Quizzes_Module1Part2.pdf (285.356 KB)
Quizzes_Module1Part1.pdf (76.416 KB)

This section will host all Questions, Answers and Explanations to the Quizzes embedded within the LAMS Sequences. Some of these will also be covered in the Review Lectures.

Once you are done with the LAMS every week, you may feel free to check the solutions. The material in this section will also help you revise the LAMS quizzes in one go. Of course, you are advised to go through the LAMS first, try the quizzes on your own, and then check the solutions. Hence, the solutions will only be posted for the LAMS sequences that you are supposed to complete each week.

Exercise 3- Graded exercise posted

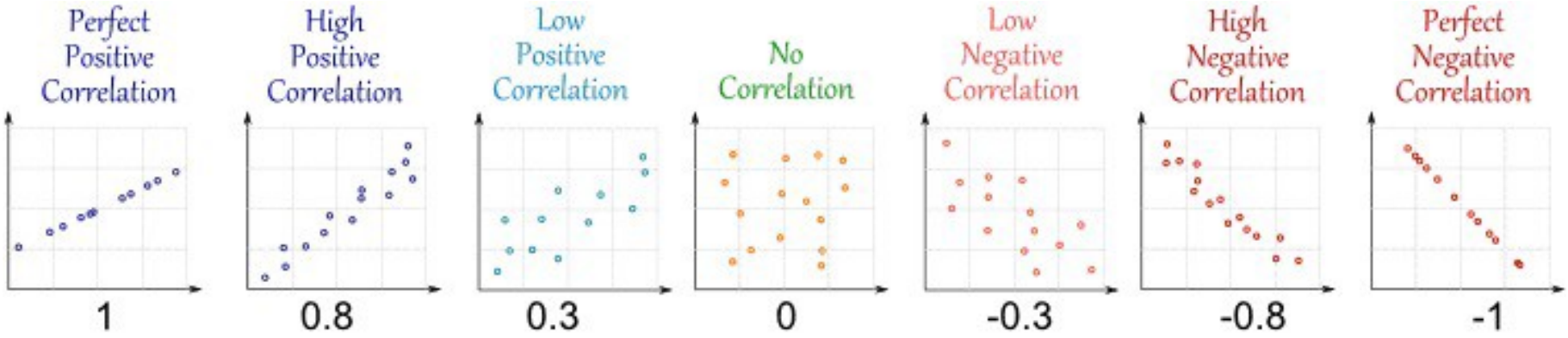
Exercise 3 : Exploratory Analysis

Attached Files: train.csv (449.879 KB)
data_description.txt (13.057 KB)
SeaBorn_CheatSheet.pdf (592.085 KB)
Exercise3_Problem.pdf (86.776 KB)

Submission is required for this Exercise. It is worth 5% of your grade.

Please check with your Lab Instructor or TA on how to submit your Solution. They will guide you with the submission of the assignment in your Coursesite for SC1015 lab_"LABGROUP"
Try to complete it during your own Lab session and submit the same before the lab session is over. Do note that the deadline for the submission is within 2 hrs from the end time of your lab group session. (HARD deadline)

You are **allowed to submit the graded lab excersise only when you are present in the lab session**. If you are absent for the graded lab excersise due to a valid reason (MC or letter of permission from school approving leave), then report to your TA and they will allow you to submit via email.



Pearson Correlation Coefficient is only defined for Numeric vs Numeric

What about the other variables?

Numeric vs Ordered Categorical
Numeric vs Unordered Categorical
Ordered Categorical vs Ordered Categorical
Unordered Categorical vs Unordered Categorical