BIKE SHARING ASSIGNMENT QUESTIONS AND ANSWERS (Shasheesh Rane)AI & ML - C36:

Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

This is taken from the Initial Inferences calculated above in the Data Preparation and EDA

- Most Booking in Season type Fall. And, in each season the Booking count has increased drastically from 2018 to 2019.
- Most of the bookings have been done during the month of may, june, july, aug, sep and oct. bookingTrend increased starting of the year till mid of the year and then it started decreasing as we approached the end of year.
- Clear weather attracted more booking Due to better Visibility in Decreasing order of Clear, Misty . Light_snowrain
- Wed, Thu, Fri, Sat have more number of bookings as compared to the other weekdays.
- When it's not holiday, booking seems to be less in number which seems reasonable as on holidays, people may want to spend time at home and enjoy with family. Time can be allocated for Outings and physical activity like cycling/bikes.
- Booking seemed to be almost equal either on working day or non-working day.

  2019 attracted more number of booking from the previous year, which shows good progress in terms of business.

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

- During the Process of Creating dummy variables , extra column is already present e.g drop_first = True is important to use, as it helps in reducing the extra column created during dummy variable creation. It helps reduce the correlations created among dummy variables.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

- The Variable 'temp' has the highest correlation with the target variable

4. How did you validate the assumptions of Linear Regression after building the model on the
training set? (3 marks)

I validated the assumptions of the Linear Regression using the 3 Properties as below :

- **ERROR TERMS NORMALITY** - The Distribution of error terms with normal evident from histogram
- **LINEAR RELATIONSHIP** - From the scatter plot linear relationship was visible with the variables
- **MULTI-COLINEARITY CHECK** - For the Linear model, the heat map showed no multi-collinearity with the variable

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

The Below Variables are totally based on the calculations above: from round(lr_6.params,3)

- **temp** = 0.549892,
- **year** = 0.233139,
- **winter** = 0.130655,

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

- Introduction to Linear Regression

Linear regression may be defined as the statistical model that analyzes the linear relationship between a dependent variable with a given set of independent variables. Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).

- Mathematically the relationship can be represented with the help of following equation −

$$Y=mX+b$$

Here, Y is the dependent variable we are trying to predict.

X is the independent variable we are using to make predictions.

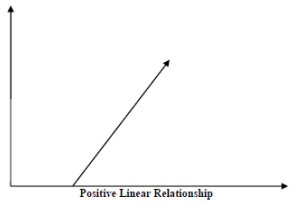
m is the slope of the regression line which represents the effect X has on Y

b is a constant, known as the YY-intercept. If X = 0,Y would be equal to bb.

Furthermore, the linear relationship can be positive or negative in nature as explained below −

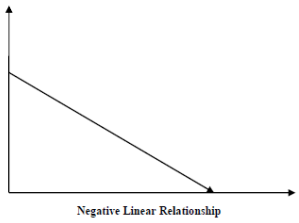
- **Positive Linear Relationship**

A linear relationship will be called positive if both independent and dependent variables increase. It can be understood with the help of following graph −



- Negative Linear relationship

A linear relationship will be called positive if independent increases and dependent variable decreases. It can be understood with the help of following graph −



Types of Linear Regression

Linear regression is of the following two types −

- Simple Linear Regression
- Multiple Linear Regression

Assumptions

The following are some assumptions about dataset that is made by Linear Regression model −

Multicollinearity − Linear regression model assumes that there is very little or no multicollinearity in the data. Basically, multicollinearity occurs when the independent variables or features have dependency in them.

Auto-correlation − Another assumption Linear regression model assumes is that there is very little or no auto-correlation in the data. Basically, auto-correlation occurs when there is dependency between residual errors.

Relationship between variables − Linear regression model assumes that the relationship between response and feature variables must be linear.

---

2. Explain the Anscombe's quartet in detail. (3 marks)

- My Understanding of the Anscobe Quartet is used before Plotting the Data. Even if the mean , Variance, correlation looks same.
- The Anscombe Quartet gives and idea how different the Data Sets are:

It is used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets

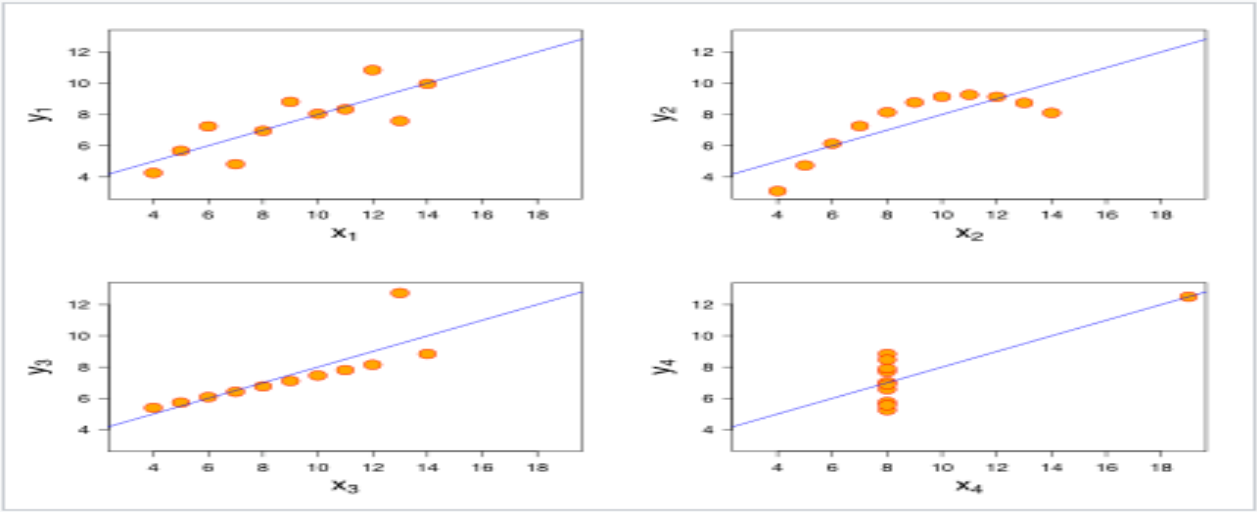
- Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed.

Each dataset consists of eleven (x,y) points.
They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analyzing it, and the effect of outliers and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough.

Reference diagrams from Wikipedia:

Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x : s_x^2	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y : s_y^2	4.125	±0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression : R^2	0.67	to 2 decimal places

A Classic Example of Anscombe Quartet that shows the Difference in the Plots of all 4 DataSets below:

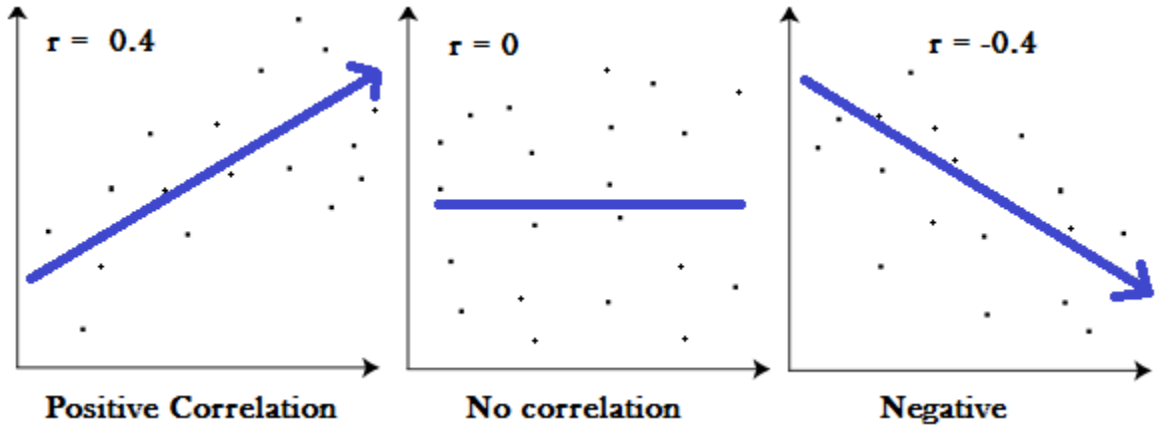


All four sets are identical when examined using simple summary statistics, but vary considerably when graphed

3. What is Pearson's R? (3 marks)

- Pearsons R Also Known as Pearson's relationship Co-efficient , value measures the strength of the Linear relationship between 2 Variables.
- The Pearsons R lies between -1 < Pearson R < 1

- It was developed by Karl Pearson from a related idea introduced by Francis Galton in the 1880s, and for which the mathematical formula was derived and published by Auguste Bravais in 1844 (



$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient
x_i = values of the x-variable in a sample
$\bar{x}$ = mean of the values of the x-variable
y_i = values of the y-variable in a sample
$\bar{y}$ = mean of the values of the y-variable

Two other formulae are commonly used: the sample correlation coefficient and the population correlation coefficient.

Sample correlation coefficient:

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

S_x and s_y are the sample standard deviation, and s_{xy} is the sample covariance

Population correlation coefficient:

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

The population correlation coefficient uses σ_x and σ_y as the population standard deviations, and σ_{xy} as the population covariance.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

- It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.
- Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.
  It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Normalization/Min-Max Scaling:
It brings all of the data in the range of 0 and 1. sklearn.preprocessing.
MinMaxScaler helps to implement normalization in python.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

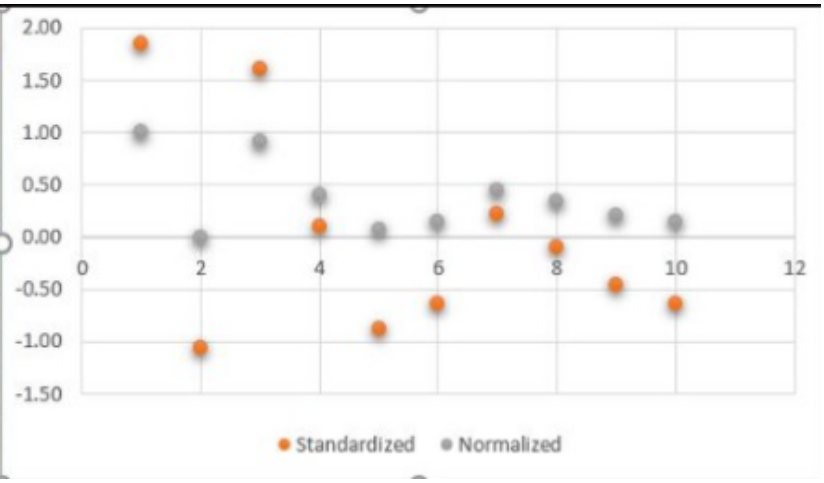
Standardization Scaling:
Standardization replaces the values by their Z scores.
It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

sklearn.preprocessing.scale helps to implement standardization in python.
One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

EXAMPLE REFERENCE: Wikipedia

Original	Standardized	Normalized
59	1.85	1.000
11	-1.06	0.000
55	1.61	0.917
30	0.09	0.396
14	-0.88	0.063
18	-0.64	0.146
32	0.21	0.438
27	-0.09	0.333
21	-0.46	0.208
18	-0.64	0.146



The Differences between Normalization and Standardized scaling is as below:

S.NO.	Normalization	Standardization
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called `MinMaxScaler` for Normalization.	Scikit-Learn provides a transformer called `StandardScaler` for standardization.
6.	This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
7.	It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
8.	It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization.

Ref: https://www.geeksforgeeks.org/normalization-vs-standardization/

My Take on MAJOR Difference in Normalized and Standardized is the :

- **The Normalized is affected by Outliers more than standardized.**
- **The Normalized is in a fixed Range of [-1,1].**
- **The Normalized used min/max & Standardization uses mean/standard dev for scaling.**

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?(3 marks)

- **If all the independent variables are orthogonal to each other, then VIF = 1.0. If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables.**

WHY?:

- **In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity**

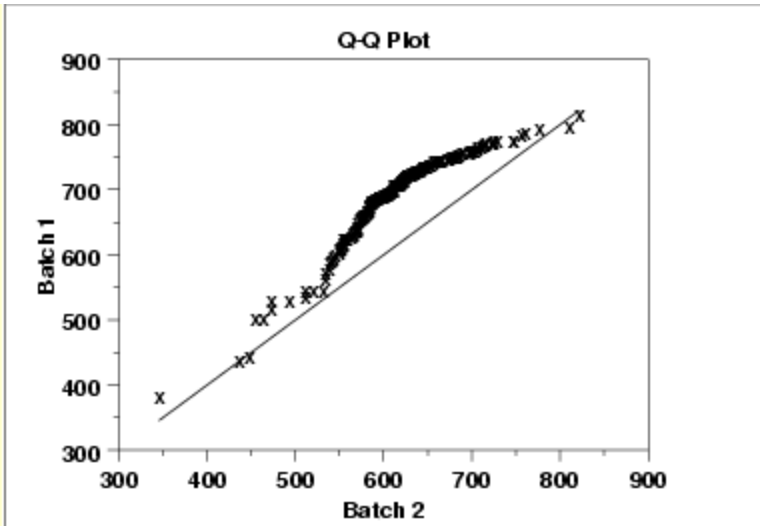
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

- The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.
  A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.
- By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value.

- A 45-degree reference line is also plotted.
- If the two sets come from a population with the same distribution,
  the points should fall approximately along this reference line.
  The greater the departure from this reference line,
  The greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

REF: https://www.itl.nist.gov/div898/handbook/eda/section3/qqplot.htm

EXAMPLE Q-QPLOT BELOW:



Q-Q plots are available in some general purpose statistical software programs.
If the number of data points in the two samples are equal,
it should be relatively easy to write a macro in statistical programs that do not support the q-q plot.
If the number of points are not equal, writing a macro for a q-q plot may be difficult.