Submitted by: Harsha Vardhan Tamma

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?

Ans: *The final Multiple Regression model contains many predictor variables that are categorical, which effect the dependent variable. After analysis from the visualization, we can infer below points:*

- *Fall season seems to have more bookings compared to other seasons. And bookings in each season increased drastically in 2019 when compared to 2018.*
- *Most of the bookings have been done during the months of May, June, July, August, September and October. There is an increasing trend during the beginning of the year and decreasing trend towards the end of the year.*
- *Clear weather (with or without few clouds) attracted more bookings when compared to other weather situations.*
- *Rentals are more in working days when compared to holidays (as people might be stay at home or visit other places with family during holidays).*
- *Bookings are almost equal either on working or non-working days.*
- *Thursday, Friday, Saturday and Sunday have more number of bookings as compared to the start of the week.*
- *2019 has more bookings when compared to 2018, which shows good progress in the business.*

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

Ans: *During creation of dummy variables,* ***drop_first = True,*** *helps in reducing extra column and reduces the correlations created among the dummy variables. This makes the model less complex.*

Syntax - ***drop_first: bool****,* ***default False****, which implies whether to get k-1 dummies out of k categorical levels by removing the first level.*

For example:
Let's say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If value A is 1 then value of B & C is 0, if value B is 1 then value of A & C is 0. Therefore if the value of A & B is 0 then definitely it would be C . So we don't need three variables.

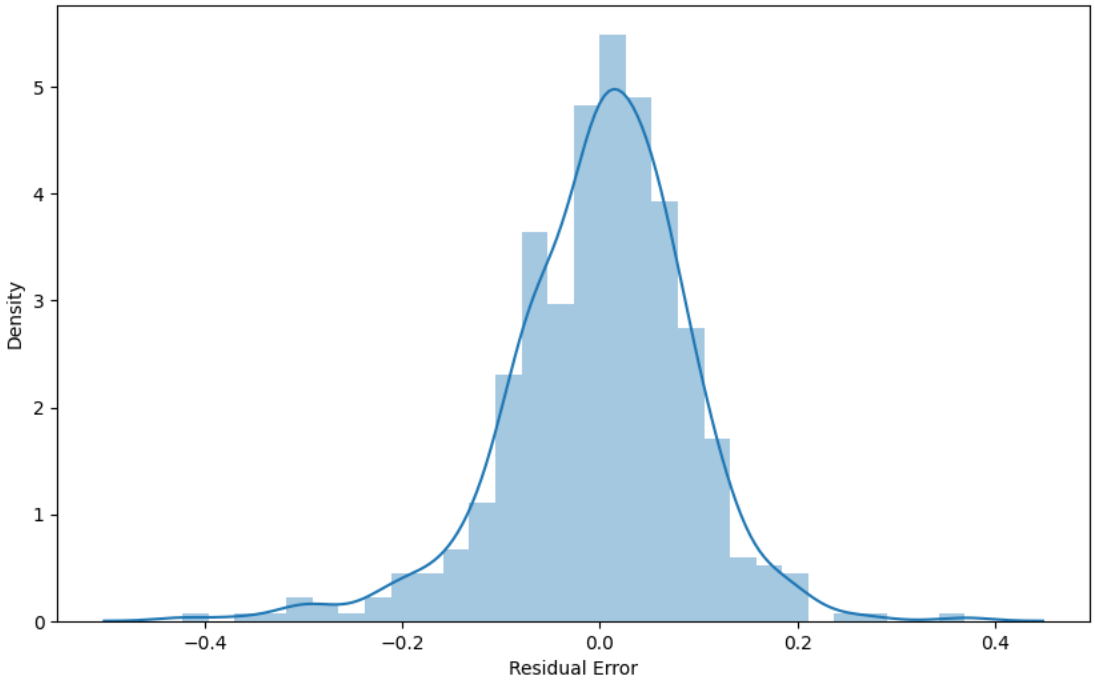
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans: *'****temp'*** *variable has the highest correlation with the target variable. As per the correlation heatmap, correlation coefficient between* ***temp*** *and* ***cnt*** *is 0.63.*

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

Ans: *To validate assumptions of the model, we go with the following procedures:*

- ***Residual Analysis:***
  We need to check if the error terms are also normally distributed. I have plotted the histogram of the error terms and this is what it looks like:



  The residuals are following the normal distribution with a mean 0.

- ***Linear relationship between predictor variables and target variable:***
  This is because all the predictor variables are statistically significant (p-values are less than 0.05). Also, R^2 value on training set is 0.833 and adjusted-R^2 value on training set is 0.830. This means that variance in data is being explained by all these predictor variables.

- ***Error terms are independent of each other:***
  The predictor variables are independent of each other. Multicollinearity issue is not there because VIF (Variance Inflation Factor) for all predictor variables are below 5.

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

Ans: *Top 3 features significantly contributing towards demand of shared bikes are:*

- ***temp*** *(coef: 0.4514)*
- ***year*** *(coef: 0.234)*
- ***sep*** *(coef: 0.0577)*

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans: *Linear regression may be defined as the statistical model that analyses the linear relationship between a dependent variable with 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).*

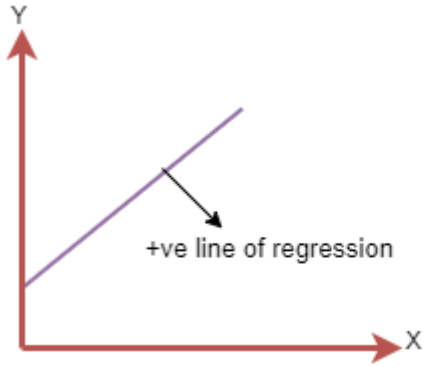
Mathematically the relationship can be represented with the help of following equation –

$$y = \beta_0 + \beta_1 x + \varepsilon$$

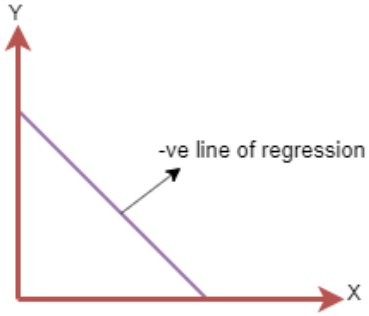
Here,
***y** is the **dependent variable** we are trying to predict (a.k.a **target variable**).*
***x** is the **independent variable** we are using to make predictions (a.k.a **predictor variable**).*
β_0 *is the **intercept** of the line. (a.k.a **constant**)*
β_1 *is the linear regression **coefficient** (scale factor to each input value, a.k.a **slope**)*
ε *is the **random error** component (a.k.a **random error term**)*

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 variable increases. 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*



Linear regression can be further divided into two types:

- ***Simple Linear Regression***

If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.

- ***Multiple Linear Regression***

If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

Assumptions:

Below are some important assumptions of Linear Regression. These are some formal checks while building a Linear Regression model, which ensures to get the best possible result from the given dataset.

- ***Linear relationship between the features and target:***
  Linear regression assumes the linear relationship between the dependent and independent variables.
- ***Small or no multicollinearity between the features:***
  Multicollinearity means high-correlation between the independent variables. Due to multicollinearity, it may difficult to find the true relationship between the predictors and target variables. Or we can say, it is difficult to determine which predictor variable is affecting the target variable and which is not. So, the model assumes either little or no multicollinearity between the features or independent variables.
- ***Homoscedasticity Assumption:***
  Homoscedasticity is a situation when the error term is the same for all the values of independent variables. With homoscedasticity, there should be no clear pattern distribution of data in the scatter plot.
- ***Normal distribution of error terms:***
  Linear regression assumes that the error term should follow the normal distribution pattern. If error terms are not normally distributed, then confidence intervals will become either too wide or too narrow, which may cause difficulties in finding coefficients.
  It can be checked using the ***q-q plot****. If the plot shows a straight line without any deviation, which means the error is normally distributed.*
- ***No autocorrelations:***
  The linear regression model assumes no autocorrelation in error terms. If there will be any correlation in the error term, then it will drastically reduce the accuracy of the model. Autocorrelation usually occurs if there is a dependency between residual errors.

2. Explain the Anscombe's quartet in detail.

Ans: *Anscombe's Quartet is the modal example to demonstrate the importance of data visualization which was developed by the statistician Francis Anscombe in 1973 to signify both the importance of plotting data before analysing it with statistical properties.*

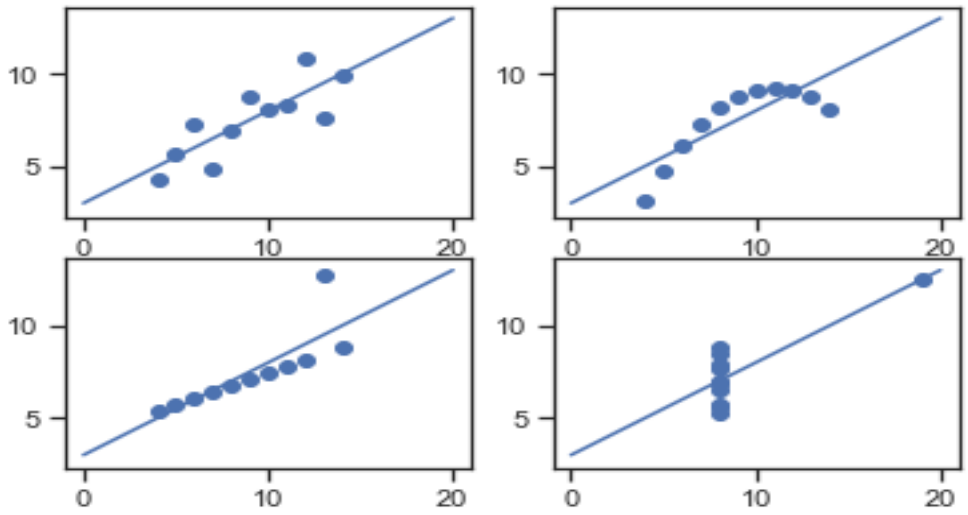
It comprises of four data-set and each data-set consists of eleven (x,y) points. The basic thing to analyse about these data-sets is that they all share the same descriptive statistics(mean, variance, standard deviation etc) but different graphical representation. Each graph plot shows the different behaviour irrespective of statistical analysis.

x1	y1	x2	y2	x3	y3	x4	y4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Apply the statistical formula on the above data-set,

Average Value of x = 9
Average Value of y = 7.50
Variance of x = 11
Variance of y =4.12
Correlation Coefficient = 0.816
Linear Regression Equation : y = 0.5 x + 3

However, the statistical analysis of these four data-sets are pretty much similar. But when we plot these four data-sets across the x & y coordinate plane, we get the following results & each pictorial view represent the different behaviour.



Graphical Representation of Anscombe's Quartet

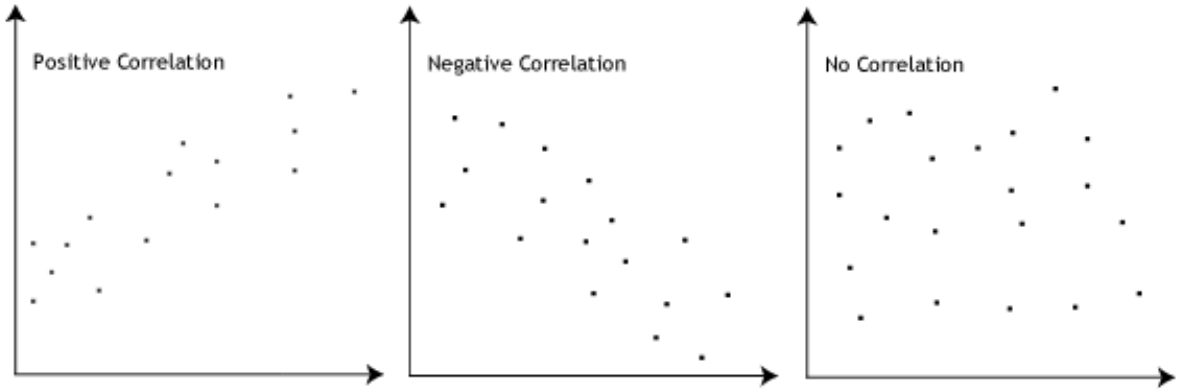
- *Data-set I — consists of a set of (x,y) points that represent a linear relationship with some variance.*
- *Data-set II — shows a curve shape but doesn't show a linear relationship (might be quadratic?).*
- *Data-set III — looks like a tight linear relationship between x and y, except for one large outlier.*
- *Data-set IV — looks like the value of x remains constant, except for one outlier as well.*

This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

3. What is Pearson's R?

Ans: *In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus it is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1.*
The Pearson's correlation coefficient varies between -1 and +1 where:

- *r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)*
- *r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)*
- *r = 0 means there is no linear association*
- *r > 0 < 5 means there is a weak association*
- *r > 5 < 8 means there is a moderate association*
- *r > 8 means there is a strong association*



Pearson's r Formula

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

Here,

- *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*

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans: *Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.*

Example: If an algorithm is not using feature scaling method, then it can consider the value 1000 meter to be greater than 5 km but that's not true and, in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to same magnitudes and thus, tackle this issue.

Normalized scaling	***Standardized scaling***
Minimum and maximum value of features are used for scaling.	*Mean and standard deviation is used for scaling.*
It is used when features are of different scales.	*It is used when we want to ensure zero mean and unit standard deviation.*
Scales values between [0, 1] or [-1, 1].	*It is not bounded to a certain range.*
It is really affected by outliers.	*It is much less affected by outliers.*
Scikit-Learn provides a transformer called MinMaxScaler for normalization.	*Scikit-Learn provides a transformer called StandardScaler for standardization*

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

Ans: *If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.*

When the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) infinity. To solve this we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

Ans: *Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. 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.*

For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q-Q plots is to find out if two sets of data come from the same distribution. A 45-degree angle is plotted on the Q-Q plot; if the two data sets come from a common distribution, the points will fall on that 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.

Importance of Q-Q plot:
When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.

statsmodels.api *provide* ***qqplot*** *and* ***qqplot_2samples*** *to plot Q-Q graph for single and two different data sets respectively.*