Assignment based Subjective Questions

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

Answer: Per my analysis, categorical variables such as, year, season, weekday, etc had a major impact/effect on our dependable variable which is "Count". There is not considerable interference by variables such as weekday and workingday.

- In year 2018 and 2019, in summer season the count of registered and casual users has increased in comparison of other seasons.
- Apparently Good Weather attracts more users.
- Compared to other weekdays, weekend has more users.
- According to the trend months with good weather have been observed to have the maximum number of users/booking.
- Considering that people in general prefer spending their time at home or with their families, there is an obvious decline in bookings/user registration during holidays.
- Comparatively to 2018, 2019 has more bookings/users proving progress for good business.
- Users/Bookings are similar on workingdays as to weekdays.

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

Answer: In order to remove the first column created for its first unique value of a column, we use drop_first=True. Considering the default value of drop_first is set to False, if we do not change it to True, it will inevitably cause to create one dummy variable each for every categorical variable in the input.

In easy words, if there was a team of 4 players playing ball and you were looking for the ball. If A, B, C players didn't have the ball, it is obvious that the 4the player, named D has it. Hence 4th dummy variable wouldn't be required to be created. That's exactly why we use drop_first=True during dummy variable creation.

Syntax :

```
pandas.get_dummies(data, prefix=None, prefix_sep='_', dummy_na=False, columns=None, sparse=False, drop_first=False, dtype=None)
```

For Example: *season_dataframe = pd.get_dummies(bike_df["season"], drop_first=True)*

In our file we dropped one of the column from season dataframe, leaving us with 3 columns (Spring, Winter, Summer).

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

Answer. "temp" and "atemp" are the variables with the highest correlation with our target variable "cnt".

```
plt.figure(figsize=(15,10))

sns.heatmap (bike_df.corr(),annot=True, cmap="Blues")

plt.show()
```

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

Answer: Assumptions for Linear Regression Model:

1. Validation for Linear Relationship: clear linear visibility amongst a few variables.
2. Normal distribution of Error Terms: Normally distributed Error Terms.
3. Multicollinearity check: Among variables there should be insignificant multicollinearity.
4. Auto Co-relation check: No independent residuals.
5. Homoscedasticity: No visible patterns in residual values.
6. No hidden or duplicate or missing variables: Include all relevant explanatory variables.

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

Answer. Top 3 features contributing significantly towards explaining the demand of the shared bikes are:

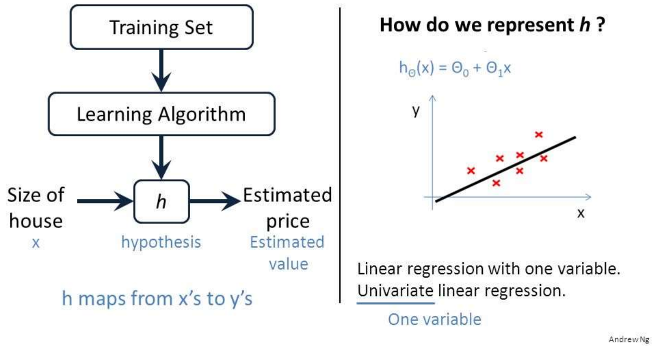
- Humidity
- Good weather
- Temp

==

General Subjective Questions

Question 1. Explain the linear regression algorithm in detail.

Answer. Linear Regression is the supervised Machine Learning model. It is used to find the best fit linear line between independent and dependent variables if a linear relationship is found. Variables on x-axis and y-axis should be linearly correlated. Based on supervised learning this is a Machine Learning algorithm, where we check the data in a continuous manner. Regression task contains Regression models, a target dependent variable based on independent variables. This is commonly used to find relationship between variables present in a model. Different regression models differ based on – the kind of relationship between dependent and independent variables, and the number of independent variables being used in the model. Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output) and regression line which is a best fit for our model.



Y = MX + C

Hypothesis function for Linear Regression

In the Training Model we are usually given as follows :

- X: input training data (univariate – one input variable(parameter))
- Y: labels to data (supervised learning) When training the model – it fits the best line to predict the value of Y for a given value of X.
- The model gets the best regression fit line by finding the best C and M values.
- C: intercept
- M: coefficient of X

Mathematical Approach:

$$\sum {e_i}^2 = \sum (Y_i - \hat{Y}_i)^2$$

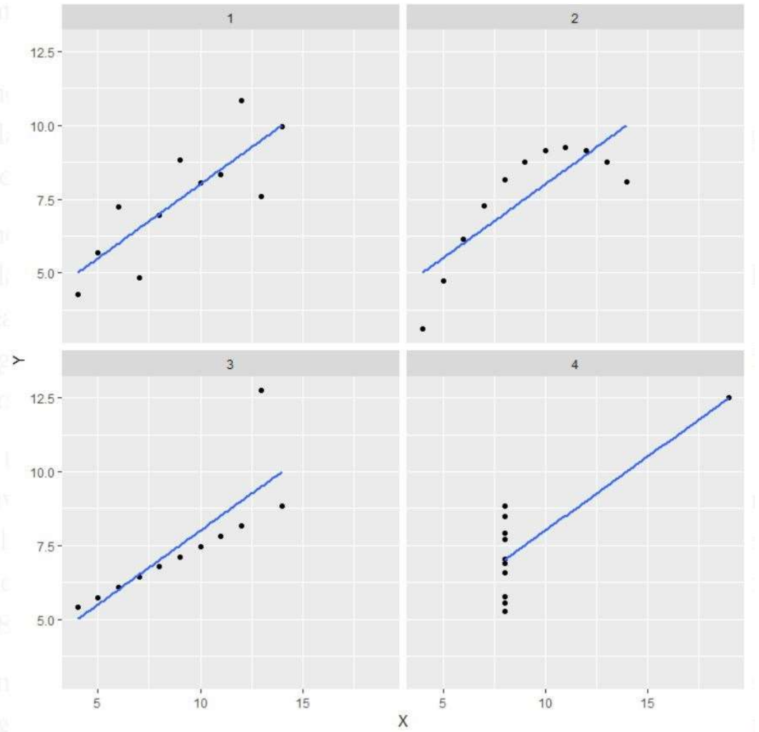
Residual/Error = Actual values – Predicted Values

Sum of Residuals/Errors = Sum(Actual- Predicted Values)

Square of Sum of Residuals/Errors = (Sum(Actual- Predicted Values))2

Question 2. Explain the Anscombe's Quartet in detail.

Answer. Anscombe's Quartet entails 4 data sets that have nearly identical descriptive statistics, but when we use charts or graphs and especially scatter plots, they have very different distributions and appearance. While working with Linear Regression Model, one can notice some peculiarities in the dataset that can manipulate one's Regression Model once built. Each dataset consists of eleven (x,y) points.



ANSCOMBE'S QUARTET FOUR DATASETS

- **1st Data Set:** Fits the Linear Regression Model perfectly.
- **2nd Data Set:** Non-Linear data causes a misfit with the Linear Regression Model.
- **3rd Data Set:** Outliers can be observed, but cannot be handled by the Linear Regression Model.
- **4th Data Set:** Outliers can be observed again, but just like 3rd Data Set, they cannot be handled by the Linear Regression Model.

APPLICATION:

The Anscombe's Quartet is often used to illustrate the importance of looking at a set of data graphically before starting to analyse according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

Question 3. What is Pearson's R?

Answer. The Pearson Correlation Coefficient (r) is the most common way of measuring a Linear Correlation. It is a number between –1 and 1 to measure the strength and direction of the relationship between two variables.

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

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

We can calculate a correlation between two numerical values or between two categorical values using Pearson's R. However, an analyst can do the same between various types of variables. One method to calculate the correlation of a numerical variable with a categorical value is to convert the numerical variable into categories. Pearson's R summarizes a dataset's characteristics. It helps describing the strength and direction between 2 quantitative variables for their linear relationship. It is also an inferential statistic, basically it means that one can use it to test Statistical Hypothesis.

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

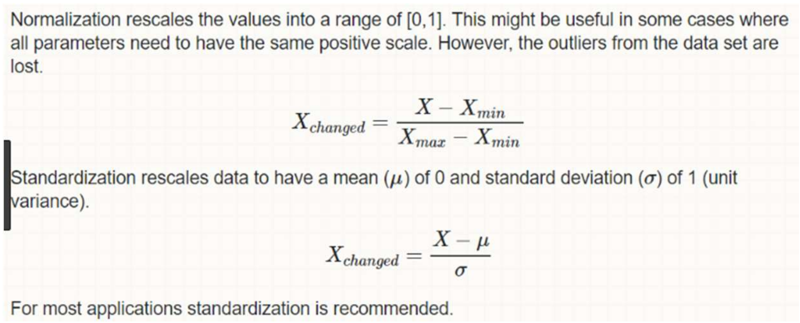
Answer. We use scaling as one of the steps through data Pre-Processing, is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm. Most of the times, collected data set contains features highly varying in magnitudes, 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.

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).
- 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.



Normalization rescales the values into a range of [0,1]. This might be useful in some cases where all parameters need to have the same positive scale. However, the outliers from the data set are lost.

$$X_{changed} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Standardization rescales data to have a mean (μ) of 0 and standard deviation (σ) of 1 (unit variance).

$$X_{changed} = \frac{X - \mu}{\sigma}$$

For most applications standardization is recommended.

S.NO.	Normalization	Standardization
1.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
2.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
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.

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

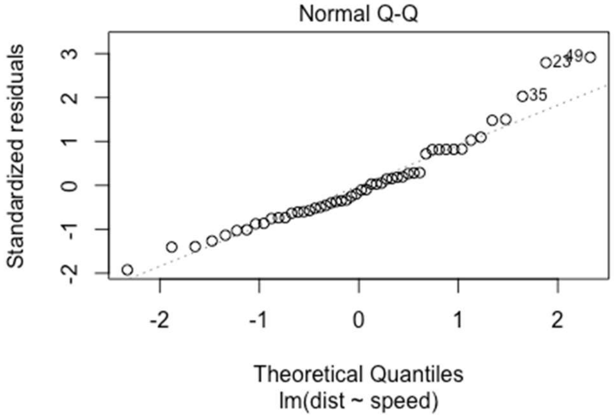
Answer. Variance inflation factor or as it is commonly known as VIF, is used to calculate the amount of multicollinearity in a set of Multiple Regression Variables. Mathematically, the VIF for a regression model variable is equal to the ratio of the overall model variance to the variance of a model that includes only that single independent variable. A high VIF indicates that the independent variable is highly collinear with other features.

A multiple regression is used when a person wants to test the effect of multiple variables on a particular outcome. An Infinite VIF happens when the dependent variable is the outcome that is being acted upon by the independent variables. Multicollinearity exists when there is a linear relationship, or correlation, between one or more of the independent variables. An infinite VIF value indicates that the corresponding variable could create a perfect linear combination with other variables. A high VIF value indicates that there is a strong correlation between variables. A general rule of thumb is that if VIF > 10 then there is multicollinearity.

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

Answer. Quantile-Quantile or it is generally known Q-Q plot, is a graphical asset to help us analyse if a set of data possibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps figure out, if two data sets come from populations with a common distribution. This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot if both the data sets are from populations with same distributions. Advantages of Q-Q plot:

- It can be used with sample sizes also
- Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.



Interpretation:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

Below are the possible interpretations for two data sets.

- Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis
- Y-values < X-values: If y-quantiles are lower than the x-quantiles.
- X-values < Y-values: If x-quantiles are lower than the y-quantiles.
- Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis