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?

Answer: There are several Inferences that i could able to make after going through analysis of categorical variables:

- Fall has the highest demand in all seasons.
- Growth has been increased from 2018 to 2019, which means demand is increasing year by year.
- In Month wise comparison, we have seen that Sept month is having highest demand.
- Demand increasing start of the year to Sept, then reduced from Sept to the end of the year
- In comparison of weathersit, 'clear' weather has higher demand than 'mist' and 'hum'
- Demand in holidays is higher than rest of the days.
- Plots of Weekday and workingday not helping us to infer anything.

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

Answer: drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

If we do not drop one of the dummy variables created from a categorical variable, then it becomes redundant with dataset as we will have constant variable(intercept) which will create multicollinearity issue.

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

Answer: Temp Variable is having 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?

Answer: I have checked the following assumptions:

- Error terms are normally distributed with mean 0.
- Error Terms do not follow any pattern.
- Multicollinearity check using VIF(s).
- Linearity Check.
- Ensured the overfitting by looking the R2 value and Adjusted R2.

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

Answer: As per our final Model, the top 3 predictor variables that influences the bike booking are:

1. **Temperature (temp)** - A coefficient value of '0.373' indicated that a unit increase in temp variable increases the bike hire numbers by 0.373 units.

2. **Lightsnow** (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds) - A coefficient value of '-0.287' indicated that, w.r.t lightsnow, a unit increase in Weathersit_3 variable decreases the bike hire numbers by 0.287 units.

3. **Year (yr)** - A coefficient value of '0.235' indicated that a unit increase in yr variable increases the bike hires numbers by 0.235 units.

General Subjective Questions

1.Explain the linear regression algorithm in detail.

Answer: Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting.

There are two types of regression–

A. Simple linear regression: Model with one independent variable.
B. Multiple linear regression: Model with more than one independent variable.

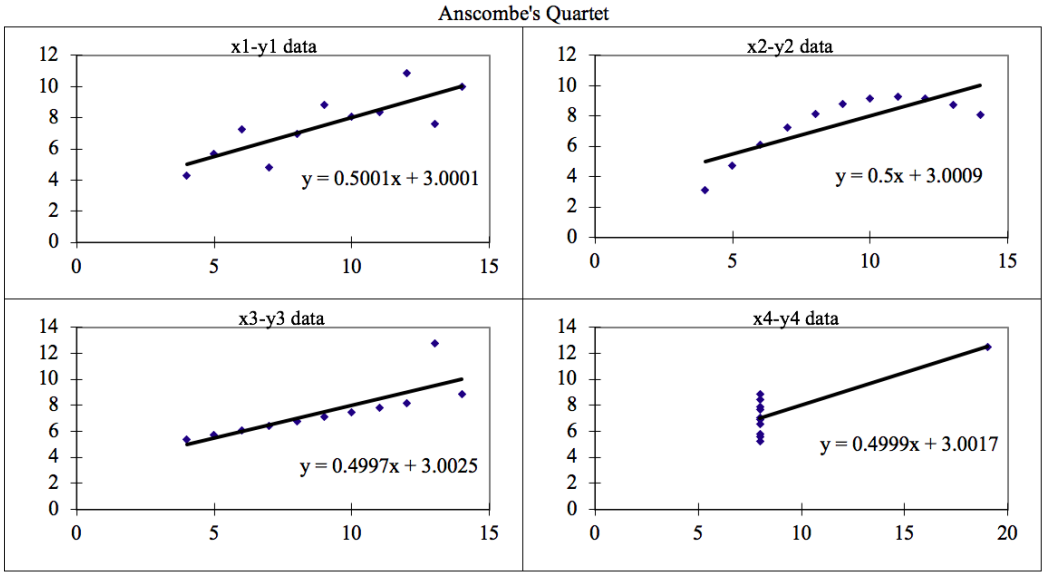
Steps that we take while building a model:

1. Reading and understanding the data.
2. Visualizing the data: If there is some obvious multicollinearity going on, this is the first place to catch it. This is where you'll also identify if some predictors directly have a strong association with the outcome variable.
3. Data preparation:
    - You can see if your dataset has columns with values as 'Yes' or 'No'. To fit a regression line, we would need numerical values and not string. Hence, we need to convert them to 1s and 0s, where 1 is a 'Yes' and 0 is a 'No'.
    - Also convert the categorical variables into numerical using dummy variables.
    - Treating the outliers if observed.
    - Treating the missing values if observed.
4. Splitting the data into train and test set.
5. Rescaling the data, if required, using the minmax scaling or standardization.
6. Dividing the train data into X and y sets for model building.
7. Building a linear model: Fit a regression line through the training data using statsmodels if statistics is of importance or else sklearn can also be used.
8. Add/remove variables unless the model has all variables with p values, VIF, r-square and prob(F-statistics) in acceptable range.
9. Residual analysis of the train data: check if the error terms are also normally distributed and also, other assumptions of linear regression.
10. Making Predictions using the final model.
11. Evaluating the model

2.Explain the Anscombe's quartet in detail.

Answer: Anscombe's Quartet can be **defined** as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots. Each dataset consists of eleven (x,y) points.

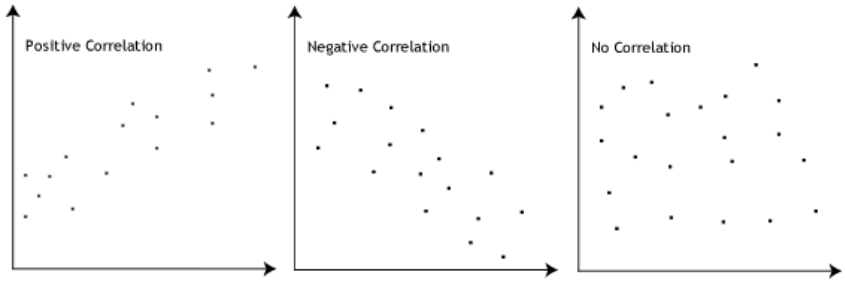
The four datasets can be described as:
Dataset 1: this **fits** the linear regression model pretty well.
Dataset 2: this **could not fit** linear regression model on the data quite well as the data is non-linear.
Dataset 3: shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model
Dataset 4: shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model



3. What is Pearson's R?

Answer: Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive.

Pearson's r measures the strength of the linear relationship between two variables.
Pearson's r always between -1 and 1.
If data lie on a perfect straight line with negative slope, then r = -1.
Positive correlation indicates the both the variable increase and decrease together. Negative correlation indicates the one the variable increase and the other variable decrease and vice versa.



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

Answer: VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. In order to determine VIF, we fit a regression model between the independent variables.

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.

The standard error of the coefficient determines the confidence interval of the model coefficients. If the standard error is large, then the confidence intervals may be large, and the model coefficient may come out to be non-significant due to the presence of multicollinearity. A general rule of thumb is that if VIF> 10 then there is multicollinearity.

An infinite VIF value indicates that the dependent variable may be expressed exactly by a linear combination of other variables. VIF = 1/ (1-R2), when R2=1 then VIF = Infinity

Example: In our Assignment, Registered Users + Casual Users = Total no. of Users If we fit the model including these 2 variables then VIF will be infinity because of this.

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

Answer: *The variance inflation factor* (VIF) *quantifies the extent of correlation between one predictor and the other predictors in a model. It is used for diagnosing* collinearity/multicollinearity. **formula for VIF=1/(1-R^2)**

If VIF is equal to infinity then it means that residual is 1 and hence it shows that there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables.

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

Answer: Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine 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 that both the data sets are from populations with same distributions.

Few advantages:

a) It can be used with sample sizes also

b) 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.

It is used to check following scenarios: If two data sets.

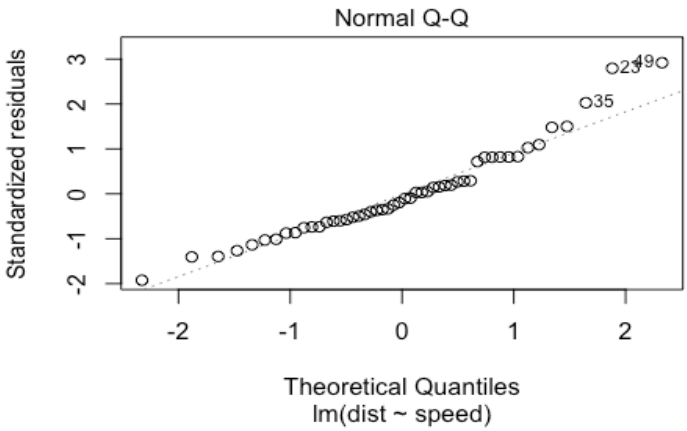
i. come from populations with a common distribution

ii. have common location and scale

iii. have similar distributional shapes

iv. have similar tail behavior

A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight i.e.


Normal Q-Q
Standardized residuals
-2
-1
0
1
2
3
23
49
35
-2
-1
0
1
2
Theoretical Quantiles
lm(dist ~ speed)