Assignment-based Subjective Questions

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans: Here are the inferences which I made from my analysis of the Categorical variables on the dependent variables i.e count are:

1) Boombikes renting were more in 2019 as compared to 2018 and also the median of 2019 (i.e 6000) is more than that of 2018 (i.e 4000).
2) The median of People renting bikes is more for holidays than to non holidays but the data is more spread out for non holiday peoples, the reason may be on holidays people prefer to rent bikes and go on drive or on vacation but they can't do that on non-holidays.
3) People renting bikes on working and non-working days are pretty similar, there is slight difference in their median but it's pretty small.
4) People renting bikes on Fall season is very high so is the median so it means that the weather condition on fall season is optimal to ride bikes.
5) People renting bikes on July is high so does the median than comes September this implies that Fall season month have high median.
6) Clear weather is most optimal for bike renting as temperature is optimal and humidity is less.
7) People renting bikes on Thursday, Friday and Sunday is more as compared to rest of the days.

Q2. Why is it important to use drop_first=True during dummy variable creation?

Ans: 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 have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.

Q3) Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans: By looking at the pair-plot among the numerical variables I found that ‘temp’ variables has highest correlation (0.63) with target variable i.e ‘count’.

Q4) How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans: The assumptions of Linear Regression can be validated after building the model on the training set:

1. Linear regression states only linear relationship between dependent and independent variables. It can be validated by plotting a scatter plot between the features and the target.
2. Homoscedasticity can be validated by using scatter plot of residual values vs predicted values.
3. Multicollinearity is a state of very high inter-correlations among the independent variables and it can be validated by using Pair-plots and Heatmaps for identifying highly correlated features.
4. The error(residuals) follow a normal distribution and it can be validated by plotting a q-q plot.
5. Autocorrelation occurs when the residual errors are dependent on each other. Autocorrelation can be tested with the help of Durbin-Watson test.

Q5) Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Ans: The top 3 features contributing significantly towards the demand of share bikes are:

- Temperature (positively correlated i.e 0.63)
- Year (positively correlated i.e 0.57)
- Season spring (negatively correlated i.e -0.56)

General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

Ans: Regression is a supervised learning technique that supports finding the correlation among variables.

Types of Regression models

1. Linear Regression
2. Polynomial Regression
3. Logistics Regression

Linear Regression

Linear regression is defined as the statistical model that analyzes 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, the value of dependent variable will also change accordingly, it may increase or decrease.

Mathematically, we can write a linear regression equation as:

$$Y = mX + b$$

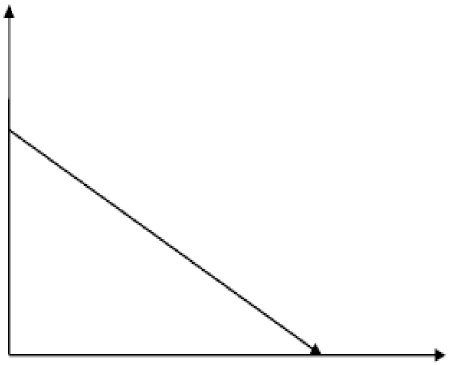
where Y is Dependent Variable, X is Independent Variable, m is the slop of the regression line which represents the effect X has on Y and b is a constant or intercept.

Positive Linear Relationship

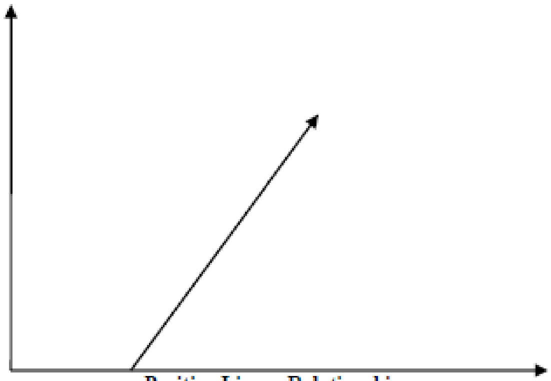
A linear relationship will be called positive if both independent and dependent variable increases. It can be understood with the help of following graph.

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.



Negative Linear Relationship



Positive Linear Relationship

There are 2 types of linear regression:

- Simple Linear Regression
- Multiple Linear Regression

Assumption of Linear Regression Model

1) Relationship between variables − Linear regression model assumes that the relationship between response and feature variables must be linear.
2) Multi-collinearity − Linear regression model assumes that there is very little or no multi-collinearity in the data. Basically, multi-collinearity occurs when the independent variables have dependency on each other.
3) Auto-correlation – Linear regression model assumes that there is very little or no auto-correlation in the data. Basically, auto-correlation occurs when there is dependency between residual errors.
4) Error term is normally distributed – Linear regression assumes that error term is normally distributed and if it is not then Central Limit Theorem is applied to error terms.
5) Error term should be Homoscedastic in nature- Linear regression model assumes that Error term is also Homoscedastic in nature it means Error term should have constant variance.
6) No Endogeneity- Linear regression assumes that there is no correlation between Error term and independent variables.

Q2) Explain the Anscombe’s quartet in detail.

Ans: Anscombe’s Quartet was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties.

It is defined as a group of four data sets which are nearly identical in simple descriptive statistics, which provides some statistical information that involves variance and mean of all x, y points in all four datasets.

It states about the importance of visualizing the data before applying various algorithms to build models which shows that the data features must be plotted in order to see the distribution of the samples that can help to identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc.

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.

Q3) What is Pearson’s R?

Ans: Correlation between sets of data is a measure of how well they are related. The most common measure of correlation in stats is the Pearson Correlation. The full name is the Pearson Product Moment Correlation (PPMC). It shows the linear relationship between two sets of data. It measures the strength of the relationship between two variables and their association with each other. In simple words, Pearson’s correlation coefficient calculates the effect of change in one variable when the other variable changes. Formula of Pearson’s R:

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

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

Ans: Scaling is the procedure of measuring and assigning the objects to the numbers according to the specified rules i.e. the process of locating the measured objects on the continuum, a continuous sequence of numbers to which the objects are assigned.

Why is scaling performed?

In regression, it is important to scale the features so that the predictors have a mean of 0. This makes it easier to interpret the intercept term as the expected value of Y when the predictor values are set to their means.

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 it leads to incorrect modelling. Thus, scaling is done to bring all the variables to the same level of magnitude.

What is the difference between normalized scaling and standardized scaling?

Normalization/Min-Max Scaling:

- It brings all of the data in the range of 0 and1.

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

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).

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

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

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

Ans: Quantile-Quantile (Q-Q) plot, is a graphical tool which is used to assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. It helps to determine if two data sets come from populations with a common distribution.

In linear regression we use Q-Q plot to confirm whether the training and test data sets that we got separately are from populations with same distributions.