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?

Some inferences I made from my analysis of categorical variable from the dataset on the dependent variable count:

• Fall season has the highest median, from this we can say that Fall season is the most optimal to ride bike followed by summer season.

• In year 2019 bike rents have increased this might be due to fact that bike rentals are getting popular, and people are getting more aware of bike rentals. This is showing an uptrend in business.

• People are renting bike more on Non holidays as compared to Holidays, it infers that people choose rental bikes as mode of transport while going to office. It also infers that people prefer to spend time with family in personal vehicle on holidays.

• Overall median across all days is same but spread for Saturday and Sunday is bigger may be evident that those who have plans on Saturday rent the bike more.

• Working and non-working days have almost the same median although spread is bigger for non-working days as people might have plans and do not want to rent bikes.

• Clear weather is most optimal for bike renting, humidity is less, no snow, no mist and temperature is less.

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

- 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 variable.
- A variable with n levels can be represented by n-1 dummy variables. So, if we remove the first column then also, we can represent the data. If the value of variable from 2 to n is 0, it means that the value of 1st variable is 1

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

- Looking at the pairplot we can see that "temp" & "atemp" are highly correlated with the target variable "count".

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

• Residual Analysis on the train data.
• To check the error terms are normally distributed.
• Applying scaling on test set.
• Dividing test data into x and y.
• Creating X_test_new dataframe by dropping variables from X_test.
• Adding a constant variable.
• Making prediction.
• Finding R squared.
• Plotting y_test and y_pred to understand the spread.

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

• t statistic: Used to determine the p-value and hence, helps in determining whether the coefficient is significant or not

• F statistic: Used to assess whether the overall model fit is significant or not. Generally, the higher the value of F statistic, the more significant a model turns out to be

• R-squared: After it has been concluded that the model fit is significant, the R-squared value tells the extent of the fit, i.e. how well the straight line describes the variance in the data. Its value ranges from 0 to 1, with the value 1 being the best fit and the value 0 showcasing the worst.

General Subjective Questions

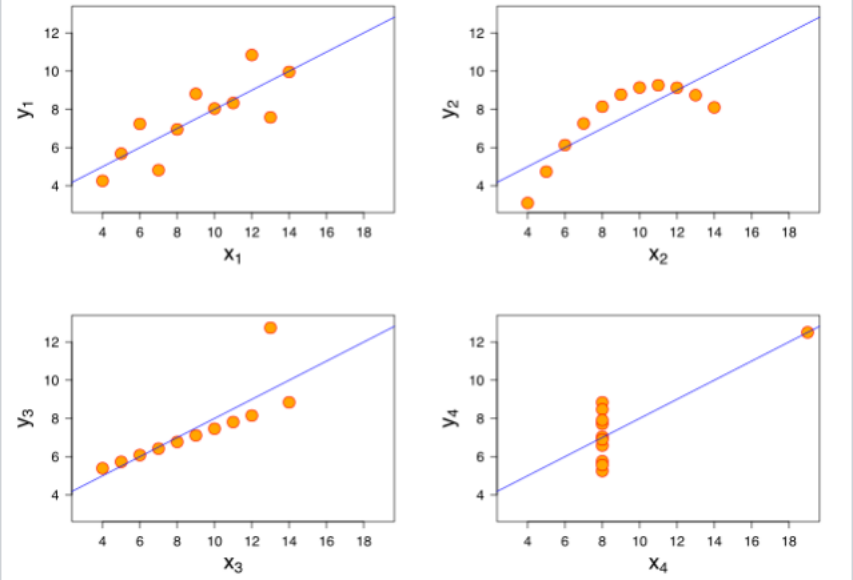
1. Explain the linear regression algorithm in detail.

• Linear regression is a quiet and simple statistical regression method used for predictive analysis and shows the relationship between the continuous variables.
• It shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), consequently called linear regression.
• If there is a single input variable (x), such linear regression is called simple linear regression.
• If there is more than one input variable, such linear regression is called multiple linear regression.
• The linear regression model gives a sloped straight line describing the relationship within the variables.
• To calculate best-fit line linear regression uses a traditional slope-intercept form (y = a0 + a1.x), where y = dependent variable, x = independent variable, a0 = intercept of line, a1 = linear regression coefficient.
• Positive Linear Relationship - If the dependent variable expands on the Y-axis and the independent variable progress on X-axis, then such a relationship is termed a Positive linear relationship.
• Negative Linear Relationship - If the dependent variable decreases on the Y-axis and the independent variable increases on the X-axis, such a relationship is called a negative linear relationship.
• 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.

2. Explain the Anscombe’s quartet in detail.

• Anscombe's quartet was developed by statistician Francis Anscombe.
• It shows us the importance of data visualization.
• It comprises of four dataset's each containing 11 (x,y) pairs, that have nearly identical simple statistical properties.
• The four dataset's appear to be similar when using typical summary statistics.
• All datasets shows the same regression line.
• Each dataset is telling the different story.

• Dataset 1 consists of a set of point that appear to follow a rough linear relationship.
• Dataset 2 consists of a neat curve but doesn't follow a linear relationship.
• Dataset 3 looks like a good relationship between x and y, except for one large outlier.
• Dataset 4 looks like x remains constant, except for one outlier.
• It was basically constructed to demonstrate both the importance of graphing data before analysing it and the effect of outliers on statistical properties.



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

3. What is Pearson's R?

• Pearson's R measures the strength of the linear relationship between two variables.
• Pearson's R is always between -1 and 1.
• When R = 1, it means if x increases y also increase, similarly if x decreases y also decrease.
• When R = -1, it means if x increases y decrease, if x decreases y increase.
• When R = 0, it means it has no relationship between x and y.

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

• Scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data pre-processing step.
• We need to do scaling so that one significant number doesn't impact the model just because of their large magnitude.
• Normalization typically means rescales the values into a range of [0,1]. Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).

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

• Variance inflation factor (VIF) is a measure of the amount of multicollinearity in a set of multiple regression variables.

• In the case of perfect correlation between two independent variables VIF = infinity. In the case of perfect correlation we get R2 = 1, which leads to 1/(1 - R2) equal to infinity. The solution for this problem is to drop one of the variable from the dataset which is causing this perfect multicollinearity.

• Model might have learned the coefficients , this might be the reason for VIF to be infinite

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

• 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. 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 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.

• A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.