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? (3 marks)

Analysis of categorical columns was performed using boxplots and bar charts. Here are some points we can deduce from the visualization –

- The fall season seems to be attracting more bookings and has skyrocketed each season from 2018 to 2019.

- Most bookings were made in May, June, July, August, September and October. This trend picked up from the beginning of the year to the middle of the year and then started to decline towards the end of the year.

- Sunny weather brought in more bookings. This is clear. Thursday, Saturday and Sunday are more booked than the beginning of the week. It seems that the number of reservations is small on days when there are no holidays, but I think there are many people who want to spend their holidays at home and have fun with their families.

- Reservations looked much the same on business and non-business days. 2019 is showing good performance with more bookings than last year.

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

drop_first = True is important to use as it helps reduce extra columns created while creating dummy variables**.** Therefore**,** it reduces the correlations produced between dummy variables.

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

Suppose you have three values in a categorical column and you want to create a dummy variable for this column. If the variable isn't A and B, it's clearly C. Therefore, no third variable is needed to identify C.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

'temp' variable has 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? (3 marks)

I have validated the assumption of Linear Regression Model based on below 5 assumptions -

- Normality of error terms - Error terms should be normally distributed
- Multicollinearity check - There should be insignificant multicollinearity among variables.
- Linear relationship validation - Linearity should be visible among variables
- Homoscedasticity - There should be no visible pattern in residual values.
- Independence of residuals - No auto-correlation

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Below are the top 3 features that go a long way to explain the demand for shared bikes –

- Temp
- Winter
- sep

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Linear regression can be defined as a statistical model that analyzes the linear relationship between a dependent variable and a given set of independent variables.
A linear relationship between variables means that a change (increase or decrease) in the value of one or more independent variables causes a corresponding change (increase or decrease) in the value of the dependent variable.

Mathematically, the relationship can be expressed using the formula − **Y = mX + c** where

- Y is the dependent variable we are trying to predict.
- X is the independent variable used to make predictions.
- m is the slope of the regression line representing the effect of X on Y.
- c is a constant known as the y-intercept. If X = 0, Y equals c.

Assumptions - Below are some assumptions about the data set produced by the linear regression model –

- Multicollinearity – The linear regression model assumes that the data have little or no multicollinearity. Basically, multicollinearity occurs when independent variables or characteristics have dependencies.
- Autocorrelation – Another assumption made by the linear regression model is that the data have little or no autocorrelation. Basically, autocorrelation occurs when there is a dependency between residual errors.
- Relationships Between Variables – The linear regression model assumes that the relationship between the response variable and the characteristic variables must be linear.
- Normality of error terms – Error terms should be normally distributed
- Homogeneity of variances – There should be no visible pattern in the residual values.

2. Explain the Anscombe's quartet in detail. (3 marks)

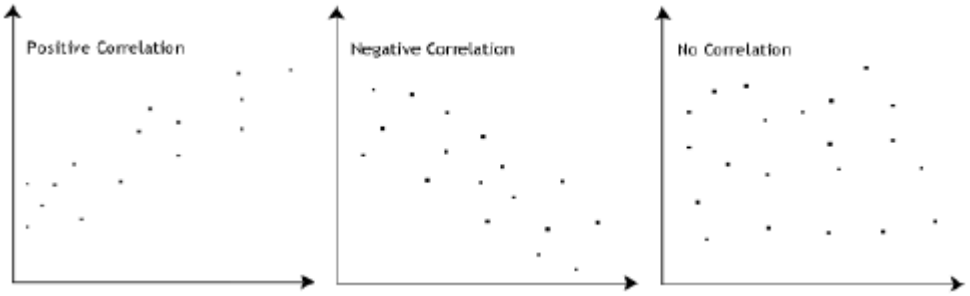
Statistician Francis Anscombe developed the Anscombe's Quartet . This contains 4 datasets, each with 11 (x,y) pairs. The most important thing to note about these datasets is that they share the same descriptive statistics. But things change completely, and when graphed they need to be completely highlighted.Each graph tells a different story, regardless of similar summary statistics.

3. What is Pearson's R? (3 marks)

Pearson's R is a numerical expression of the strength of a linear relationship between variables. The correlation coefficient is positive if the variables tend to rise and fall together. The correlation coefficient is negative if the variables tend to rise and fall as opposed to low values of one variable and high values of the other variable.

The Pearson correlation coefficient r can take values from 1 to -1. A value of 0 indicates no relationship between the two variables. Values greater than 0 indicate a positive association. That is, when the value of one variable increases, the value of the other variable also increases. Values less than 0 indicate a negative relationship. That is, as the value of one variable increases, the value of the other variable decreases.

This is shown in the diagram below



4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Feature scaling is a technique for standardizing independent features present in the data to a fixed range. This is done during data preprocessing to handle different magnitudes, values or units. When feature scaling is not performed, machine learning algorithms tend to weigh large values and consider small values to be lower, regardless of the unity of the values.

Example: If the algorithm does not use the feature scaling method, a value of 3000 meters can be considered greater than 5 km, which is actually incorrect and in this case the algorithm gives wrong predictions . So to address this issue, we use feature scaling to make all values the same magnitude.

The difference between normalized scaling and standardized scaling

1. Normalized scaling

The minimum and maximum values of the features are used for scaling. For scaling, the mean and standard deviation are used. Use when objects have different scales.
Scale values between [0, 1] or [-1, 1]. It is not limited to a specific range. It is heavily influenced by exhaust gasses. Scikit-Learn provides a MinMaxScaler transformer for regularization.

2. Standardized scaling

Mean and standard deviation are used for scaling. It is used when you want to give a mean of 0 and standard deviation of 1. It is not limited to a specific range. They are much less affected by exhaust gasses. Scikit-Learn provides a scaler called StandardScaler for standardization.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

For perfect correlation, VIF = infinity. A large VIF value indicates a correlation between variables. A VIF of 4 indicates the presence of multicollinearity, which overestimates the variance of the model coefficients by a factor of 4 or more. When the VIF value is infinite, it indicates a perfect correlation between the two independent variables.

For perfect correlation, we get R-squared(R2) = 1, so 1/(1-R2) goes to infinity. To address this problem, one of the variables that causes this perfect multicollinearity must be removed from the data set.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

A quantile-quantile (q-q) plot is a graphical method for determining whether two data sets come from populations with a common distribution.

Using QQ Chart:

A q-q plot is a plot of the quantiles in the first data set compared to the quantiles in the second data set. A quantile is the percentage (or percentage) of points below a given value. That is, the 0.3 (or 30%) quantile is the point where 30% of the data are below this value and 70% are above this value. The baseline is also applied at a 45 degree angle. If the two sets are from populations with identical distributions, the points should lie approximately along this baseline. The greater the deviation from this control line, the more evidence you infer that the two data sets are from populations with different distributions.

The importance of QQ charts:

Given two data samples, it is often desirable to know whether the assumption of a common distribution is justified. In this case, the location and size estimator can combine the two data sets to obtain an overall location and size estimate. It is also useful to get an idea of these differences if the two samples are different. A q-q plot can provide more information about the nature of the difference than analysis methods such as the chi-square test and the 2-sample Kolmogorov-Smirnov test.