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?

A1. The following are the conclusions that may be drawn from categorical variable analysis :-

1. The season has a significant impact on the quantity of bike bookings, with spring having the fewest and fall having the most.
2. Compared to 2018, there were more bookings made overall in 2019.
3. On holidays, there are substantially fewer bookings.
4. Although there appears to be a little rise on Saturday and an even smaller decline on Friday and Monday, the number of reservations appears to be mostly consistent throughout weekdays.
5. Heavy rain conditions were never present during the sample collection, however mild rain had a significant negative impact on the number of reservations. Bookings are somewhat less when the weather is overcast than when it is clear.

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

A2. Setting drop_first = True causes get_dummies to omit the dummy variable for the variable's first category. When there are K mutually exclusive categories in a categorical variable, you only need K – 1 additional dummy variables to convey the same information. This is because if all of the current dummy variables equal 0, we know that the value for the last dummy variable should be 1.

As a result, it's usual practice to remove the dummy variable for the first level of the category variable you're encoding. Indeed, it is typically required for certain types of machine learning models. Failure to remove the extra dummy variable may cause problems with your model. Because there is an unneeded correlation between the dummy variables

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

A3. Temperature appears to have the strongest correlation with our target variable.

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

A4.

a. A scatter plot was created to assess the linear relationship between the independent and target variables.
b. The residual error distribution plot was plotted to ensure that the error terms are normally distributed with a mean of zero.
c. To validate homoscedasticity and error term independence, a scatter plot was drawn between the residual errors and the target variable.

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

A5. The top three features contributing to the demand are as follows:

a. Temperature (Higher temp results in higher demand)
b. Year (Year 2019 saw higher demand)
c. Weather situation (When it's raining, demand is low)

General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

A1. Linear Regression is a supervised learning-based machine learning technique. It runs a regression test. Based on independent variables, regression models a target prediction value. It is mostly used to determine the link between variables and predicting. Different regression models differ in the type of relationship they evaluate between dependent and independent variables, as well as the amount of independent variables utilized.

Linear regression is used to predict the value of a dependent variable (y) based on a given independent variable (x). As a result, this regression technique determines a linear relationship between x (input) and y. (output). Hence, the name linear regression.

Linear regression hypothesis function:

$y = \theta 1 + \theta 2.x$

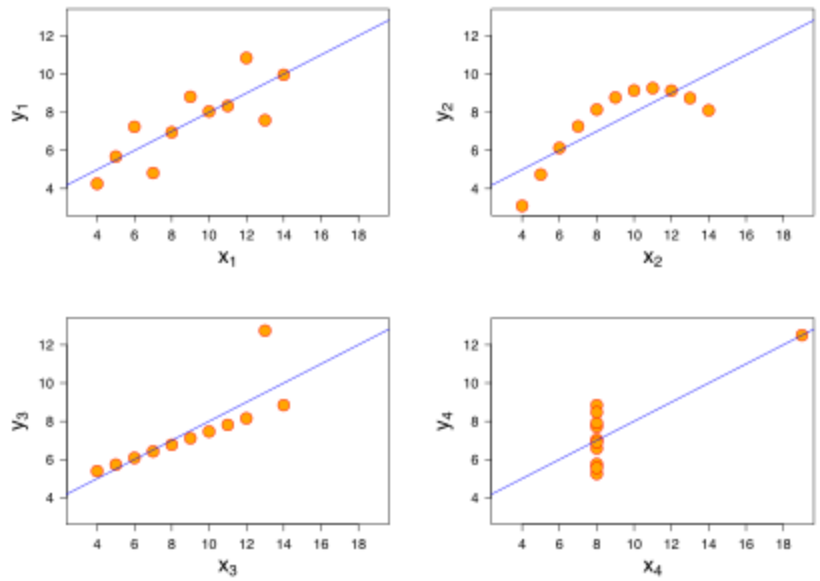
We are given the following to train the model:
x: training data input (univariate = one input variable(parameter))
y: data labels (supervised learning)

It fits the best line to predict the value of y for a given value of x when training the model. By determining the best θ1 and θ2 values, the model obtains the best regression fit line.
θ1: the intercept
θ2: the x coefficient

We get the best fit line after we get the best θ1 and θ2 values. So, when we use our model to forecast, it will predict the value of y for the input value of x.

Q2. Explain the Anscombe’s quartet in detail.

A2. Anscombe's quartet consists of four data sets with virtually similar simple descriptive statistics but radically diverse distributions and appearances when graphed. Each dataset is made up of eleven (x,y) points. They were built by statistician Francis Anscombe to highlight the significance of plotting data when evaluating it, as well as the effect of outliers and other influential observations on statistical features. These plots were created to dispel the myth among statisticians that "numerical computations are correct, while graphs are rough."



1. The first scatter plot looks to be a straightforward linear relationship, corresponding to two variables that are correlated, where y might be modeled as gaussian with a mean that is linearly dependent on x.
2. While there is a clear relationship between the two variables in the second graph, it is not linear, and the Pearson correlation coefficient is irrelevant. A broader regression and the accompanying coefficient of determination would be preferable.
3. The modeled relationship in the third graph is linear, but it should have a different regression line (a robust regression would have been called for). The estimated regression is offset by one outlier, whose influence is strong enough to reduce the correlation coefficient from 1 to 0.816.
4. Finally, the fourth graph illustrates how one high-leverage point can yield a high correlation coefficient even when the other data points show no association between the variables.

Q3. What is Pearson’s R?

A3. Correlation is a statistical measure that assesses the relationship between two variables. The correlation coefficient represents the intensity of the association between the two variables as well as the direction of the relationship. The correlation coefficient has a numerical value between -1.0 and +1.0.

A negative correlation coefficient suggests that when one variable changes, the other changes in proportion but in the opposite direction, whereas a positive correlation coefficient means that both variables change in proportion but in the same direction.

Pearson's Correlation Coefficient is also known as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation in statistics. It's a statistic that calculates the linear relationship between two variables. It, like all correlations, has a numerical value between -1.0 and +1.0.

It cannot, however, describe nonlinear interactions between two variables or distinguish between dependent and independent variables.

Pearson's correlation coefficient is calculated by dividing the covariance of the two variables by the product of their standard deviations.

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

A4. Feature scaling is a technique for standardizing the independent features present in data within a specific range. It is used during data pre-processing to deal with drastically changing magnitudes, values, or units. If feature scaling is not performed, a machine learning algorithm will tend to weight greater values as higher and consider smaller values as lower, regardless of the unit of measurement.

Min-Max Normalization rescales a feature or observation value with a distribution value between 0 and 1.
Standardization is a very effective approach for re-scaling a feature value so that it has a distribution with a mean value of 0 and a variance of 1.

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

A5. The VIF formula is as follows:

VIF = 1/1-R2

VIF approaches infinity when R2 reaches 1, i.e. when an independent variable is perfectly explained by other variables.

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

A6. The quantile-quantile (q-q) plot is a graphical tool for detecting if two data sets are from the same population.
A q-q plot is a comparison of the first data set's quantiles to the quantiles of the second data set. There is also a 45-degree reference line drawn. If the two sets are drawn from the same population, the points should fall roughly along this reference line. The larger the deviation from this reference line, the stronger the evidence that the two data sets came from populations with distinct distributions.

The q-q plot has the following advantages:

1. The sample sizes do not have to be the same.
2. Many distributional aspects can be examined at the same time. This plot, for example, can detect movements in position, shifts in scale, changes in symmetry, and the existence of outliers. For example, if the two data sets are from populations whose distributions differ solely in location, the points should lie along a straight line that is displaced up or down from the 45-degree reference line.