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)**

We can conclude the following regarding the effect of the categorical variables on the dependent variable:

a. The company should focus on expanding business during Spring.
b. The company should focus on expanding business during November & December.
c. Based on previous data it is expected to have a boom in number of users once situation comes back to normal, compared to 2019.
d. There would be less bookings during bad weather like light snow or rain.

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

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

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

The pair plots show that the numerical variables “atemp” and “temp” possess the highest correlation with the target dependent variable “cnt”.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

The following assumptions of linear progression were validated after building the model:

1. Errors are normally distributed with mean zero
2. Error terms are indepedent of one another

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

Following are the top 3 features which contribute significantly towards explaining the demand of the shared bikes:

1. atemp
2. yr
3. season_spring

General Subjective Questions:

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

Regression is a method of modelling a target value based on independent predictors. This method is mostly used for forecasting and finding out cause and effect relationship between variables. Regression techniques mostly differ based on the number of independent variables and the type of relationship between the independent and dependent variables.

Simple linear regression is a type of regression analysis where the number of independent variables is one and there is a linear relationship between the independent(x) and dependent(y) variable. The red line in the above graph is referred to as the best fit straight line. Based on the given data points, we try to plot a line that models the points the best. The line can be modelled based on the linear equation shown below:

$y=a_0+a_1 * X$

The motive of the linear regression algorithm is to find the best values for a_0 and a_1. Following are two important concepts related to linear regression:

Cost Function

The cost function helps us to figure out the best possible values for a_0 and a_1 which would provide the best fit line for the data points. Since we want the best values for a_0 and a_1, we convert this search problem into a minimization problem where we would like to minimize the error between the predicted value and the actual value.
The difference between the predicted values and ground truth measures the error difference. We square the error difference and sum over all data points and divide that value by the total number of data points. This provides the average squared error over all the data points. Therefore, this cost function is also known as the Mean Squared Error (MSE) function. Now, using this MSE function we are going to change the values of a_0 and a_1 such that the MSE value settles at the minima.

Gradient Descent

The next important concept needed to understand linear regression is gradient descent. Gradient descent is a method of updating a_0 and a_1 to reduce the cost function (MSE). The idea is that we start with some values for a_0 and a_1 and then we change these values iteratively to reduce the cost. Gradient descent helps us on how to change the values.

Linear regression models can be classified into two types depending upon the number of independent variables:

a. Simple linear regression: When the number of independent variables is 1
b. Multiple linear regression: When the number of independent variables is more than 1

The strength of a linear regression model is mainly explained by R^2, where $R^2 = 1 - (RSS / TSS)$

a. RSS: Residual Sum of Squares

b. TSS: Total Sum of Squares

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

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. It 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. There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets.

This tells us about the importance of visualizing the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. Also, the Linear Regression can be only considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets.

The four datasets can be described as:

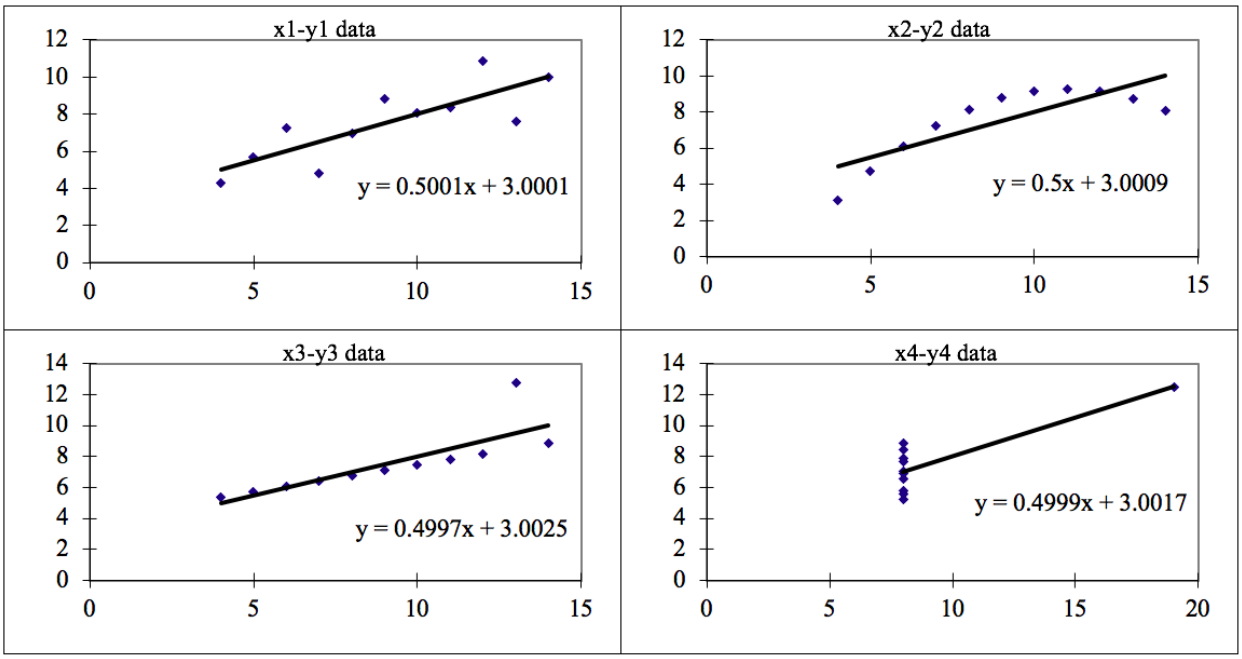
Dataset 1: this fits the linear regression model pretty well.

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

These 4 datasets can be plotted as follows:



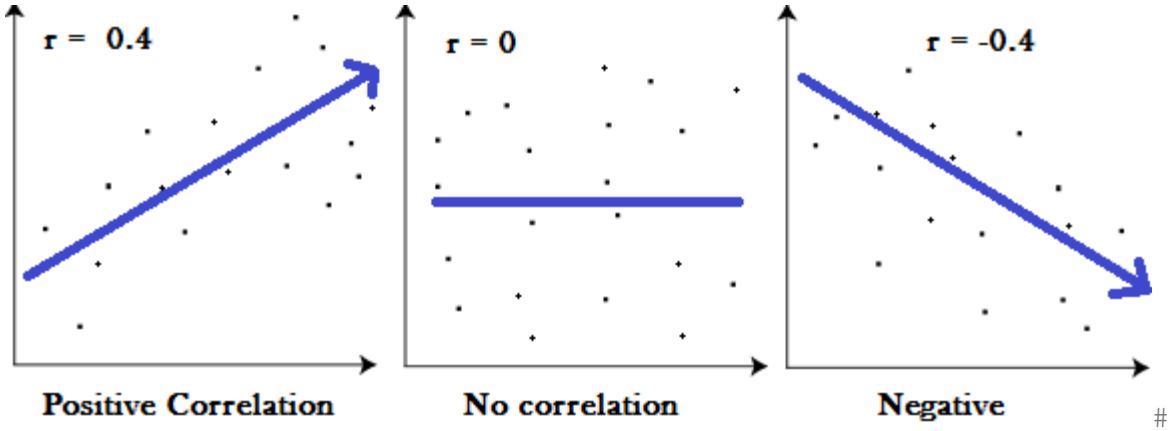
The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

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

Correlation coefficients are used to measure how strong a relationship is between two variables. There are several types of correlation coefficient, but the most popular is Pearson's Correlation (also called Pearson's R). This correlation coefficient commonly used in linear regression.

Correlation coefficient formulas are used to find how strong a relationship is between data. The formulas return a value between -1 and 1, where:

- 1 indicates a strong positive relationship.
- -1 indicates a strong negative relationship.
- A result of zero indicates no relationship at all.



#

- A correlation coefficient of 1 means that for every positive increase in one variable, there is a positive increase of a fixed proportion in the other. For example, shoe sizes go up in (almost) perfect correlation with foot length.
- A correlation coefficient of -1 means that for every positive increase in one variable, there is a negative decrease of a fixed proportion in the other. For example, the amount of gas in a tank decreases in (almost) perfect correlation with speed.
- Zero means that for every increase, there isn't a positive or negative increase. The two just aren't related.

The absolute value of the correlation coefficient gives us the relationship strength. The larger the number, the stronger the relationship. For example, |-.75| = .75, which has a stronger relationship than .65.

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.
#
#
Real Life Example
Pearson correlation is used in thousands of real life situations. For example, scientists in China wanted to know if there was a relationship between how weedy rice populations are different genetically. The goal was to find out the evolutionary potential of the rice. Pearson's correlation between the two groups was analyzed. It showed a positive Pearson Product Moment correlation of between 0.783 and 0.895 for weedy rice populations. This figure is quite high, which suggested a fairly strong relationship.

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

When there are a lot of independent variables in a model, some of them might be on very different scales which will lead a model with very weird coefficients that might be difficult to interpret. Also, the categorical variables will take either 0 or 1 as their values. Hence, it is important to have everything on the same scale for the model to be easily interpretable.

So, we need to scale features because of two reasons: 1. Ease of interpretation 2. Faster convergence for gradient descent methods. This process is called feature scaling.

The features can be scaled using two very popular methods:

1. Standardizing: The variables are scaled in such a way that their mean is zero and standard deviation is one.

2. Min-Max Scaling: The variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data.

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F statistic, p-values, R-square, etc.

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

(3 marks)

Looking at correlations might not always be useful as itis possible that just one variable might not completely explain some other variable but some of the variables combined might be able to do that. To check this sort of relations between variables, we use VIF. VIF basically helps explaining the relationship of one independent variable with all the other independent variables.

If there is perfect correlation, then VIF=Infinity.

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

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets belong to 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. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value.

A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

The advantages of the q-q plot are:

1. The sample sizes do not need to be equal.
2. Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot. For example, if the two data sets come from populations whose distributions differ only by a shift in location, the points should lie along a straight line that is displaced either up or down from the 45-degree reference line.

The q-q plot is similar to a probability plot. For a probability plot, the quantiles for one of the data samples are replaced with the quantiles of a theoretical distribution.

The q-q plot is formed by:

- Vertical axis: Estimated quantiles from data set 1
- Horizontal axis: Estimated quantiles from data set 2

Both axes are in units of their respective data sets. That is, the actual quantile level is not plotted. For a given point on the q-q plot, we know that the quantile level is the same for both points, but not what that quantile level actually is.

If the data sets have the same size, the q-q plot is essentially a plot of sorted data set 1 against sorted data set 2. If the data sets are not of equal size, the quantiles are usually picked to correspond to the sorted values from the smaller data set and then the quantiles for the larger data set are interpolated.

The q-q plot is used to answer the following questions:

- Do two data sets come from populations with a common distribution?
- Do two data sets have common location and scale?
- Do two data sets have similar distributional shapes?
- Do two data sets have similar tail behavior?

When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.