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?

   Ans:

   Analysis is done using boxplots and bar plots. Following are the conclusions:

   a. The total customer count is highest in fall season followed by summer and is lowest in spring.

   b. The customer count thus is highest in September and October ie in fall. The bookings have steadily increased in first 6 months followed by a sharp dip after October due to onset of winter season.

   c. There is not much difference in customer count throughout the week as the average customer count is almost same as seen in 3rd plot.

   d. The customer count is highest in clear weather and lowest when there is snow.

   e. The total customer count does not change much in each season for 2018 and 2019. However the total number of customers is higher in 2019

   f. The bookings are higher when it is not a holiday as when on holidays, people might be at home or may be traveling . The bookings have significantly reduced in 2019 for the same.

   g. Bookings have signnificantly increased in weekdays in 2019 with higher bookings in the second half of the week.

   h. The bookings have increased in working days in 2019.

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

   Ans:

   The drop_first= True command in dummy variable creation is needed to remove the redundant variable. It thus reduces the correlation between the dummy variables. If we have k categorical levels, we get k-1 dummy variables**.**

   For eg: If we have 3 dummy variables and we know that the value is not second or third, we know it is first, so we do not need the first dummy variable, hence we can drop it.

   Syntax:

   pandas.get_dummies(*data, drop_first=False*)

   Each variable is converted in as many 0/1 variables as there are different values

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

   Ans:

   Temp,atemp (feeling temp) and year have the highest correlation with cnt (count of total rental bikes) which is the target variable.

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

   Ans:

   a. The residuals are normally distributed and are centered around 0.
   b. The variables are linearly related.
   c. There is no visible pattern between the residuals.

5. 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 explaining the demand of the shared bikes are temp, yr and September.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

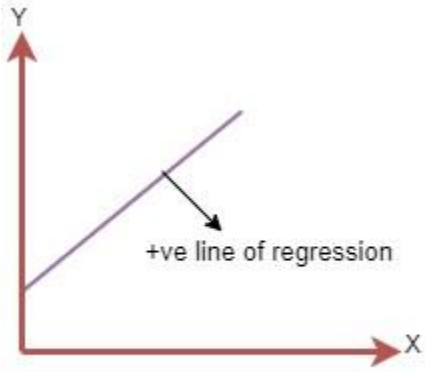
   Ans:

   Linear regression may be defined as the statistical model that analyses 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 (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).

   Mathematically the relationship can be represented with the help of following equation − Y = mX + c Here, Y is the dependent variable we are trying to predict. X is the independent variable we are using to make predictions. m is the slope of the regression line which represents the effect X has on Y c is a constant, known as the Y-intercept. If X = 0, Y would be equal to c. Furthermore, the linear relationship can be positive or negative in nature as explained below−

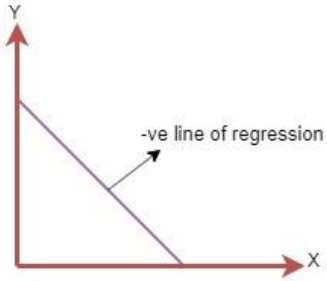
   a. Positive Linear Relationship:

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



The line equation will be: $Y= a_0+a_1x$

b. 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.



The line of equation will be: $Y= -a_0+a_1x$

c. Linear regression is of the following two types − Simple Linear Regression ¬ Multiple Linear Regression.

Assumptions - The following are some assumptions about dataset that is made by Linear Regression model –

i. 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 or features have dependency in them.

ii. Auto-correlation – Another assumption Linear regression model assumes is that there is very little or no auto-correlation in the data. Basically, auto-correlation occurs when there is dependency between residual errors.

iii. Relationship between variables – Linear regression model assumes that the relationship between response and feature variables must be linear.

iv. Normality of error terms – Error terms should be normally distributed

v. Homoscedasticity – There should be no visible pattern in residual values.

2. Explain the Anscombe's quartet in detail.

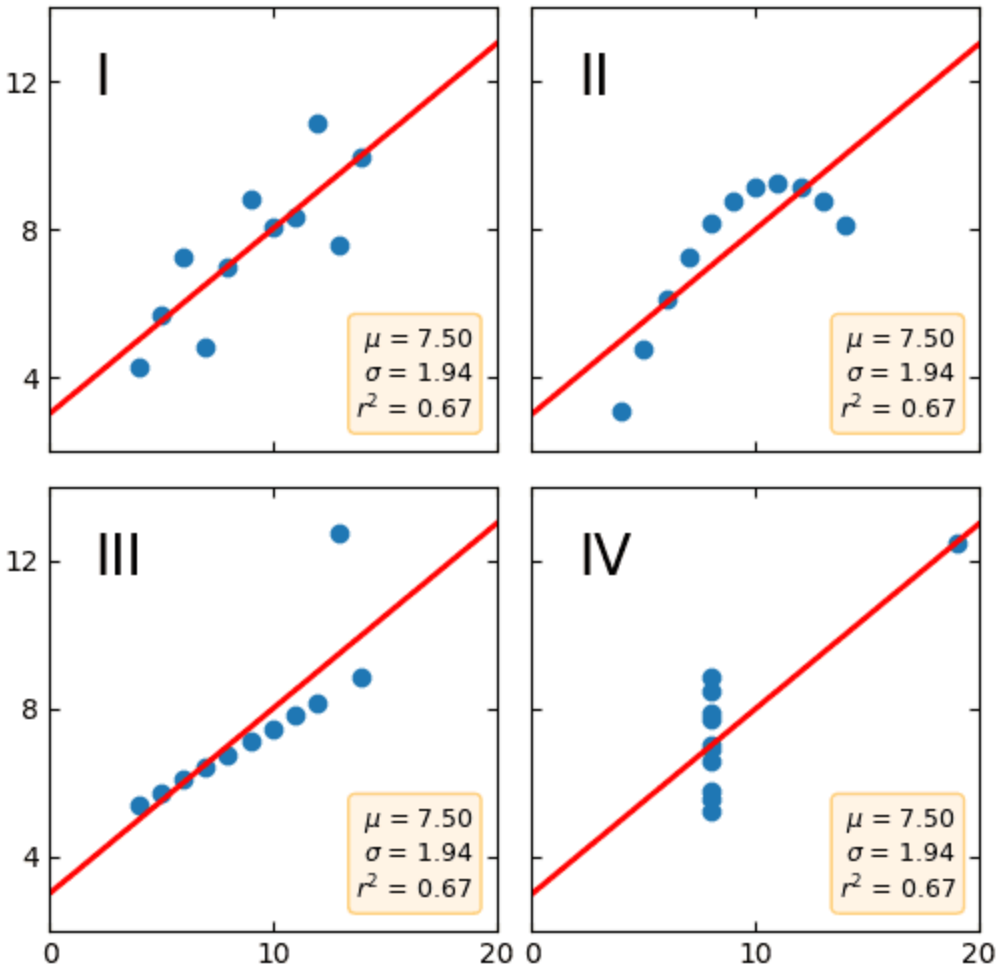
Ans:

Anscombe's quartet is a group of datasets (x, y) that have the same mean, standard deviation, and regression line, but which are qualitatively different.

It is often used to illustrate the importance of looking at a set of data graphically and

not only relying on basic statistic properties.

When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well but each dataset is telling a different story



- Dataset I appears to have clean and well-fitting linear models.
- Dataset II is not distributed normally.
- In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.
- Dataset IV shows that one outlier is enough to produce a high correlation coefficient. This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

3. What is Pearson's R?

   Ans:

   The Pearson correlation coefficient (*r*) is the most widely used correlation coefficient and is known by many names:

- Pearson's *r*
- Bivariate
- correlation
- Pearson product-moment correlation coefficient (PPMCC)
- The correlation coefficient

The Pearson correlation coefficient is a descriptive statistic, meaning that it summarizes the characteristics of a dataset. Specifically, it describes the strength and direction of the linear relationship between two quantitative variables.

Although interpretations of the relationship strength (also known as effect size) vary between disciplines, the table below gives general rules of thumb:

Pearson correlation coefficient (*r*) value	**Strength**	**Direction**
Greater than .5	Strong	Positive
Between .3 and .5	Moderate	Positive
Between 0 and .3	Weak	Positive
0	None	None
Between 0 and –.3	Weak	Negative
Between –.3 and –.5	Moderate	Negative
Less than –.5	Strong	Negative

The Pearson correlation coefficient is also an inferential statistic, meaning that it can be used to test statistical hypotheses. Specifically, we can test whether there is a significant relationship between two variables.

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

Scaling, or feature scaling, refers to the process of transforming the values of variables to a specific range. This is often done to ensure that all variables have a comparable impact on the regression model. Scaling can help prevent certain variables from dominating the model due to their larger magnitude. However, it is important to note that scaling is not a requirement for linear regression and its necessity depends on the data and the specific goals of the analysis.

Normalization	Standardized
This method scales the model using minimum and maximum values	This method scales the model using the mean and standard deviation.
When features are on various s cales, it is functional.	When a variable's mean and standard d eviation are both set to 0, it is beneficial.
Values on the scale fall betwee n [0, 1] and [-1, 1].	Values on a scale are not constrained to a particular range.
Additionally known as scaling normalization.	This process is called Z-score normalization.
When the feature distribution is unclear, it is helpful.	When the feature distribution is consistent, it is helpful.

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. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity. When the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) infinity. To solve this we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

   Ans:

   The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

   Use of Q-Q plot: A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. 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.

Importance of Q-Q plot: 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.