## Assignment-based Subjective Questions

**Question 1**. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)
**Total Marks**: 3 marks (Do not edit)
**Answer:** <Your answer for Question 1 goes below this line> (Do not edit)

From the analysis of the categorical variables from the dataset below are the inferences:

**weathersit**: The demand for shared bikes dropped drastically during snow/little rain, and it was high in clear and mist weather,this indicates that weather could be a good predictor

**season**: The demand for shared bikes was highest in fall, and lowest in spring.

**month**: The demand reached its peak during the middle of year, from June to September

**holiday**: The median of the demand is lower in holiday compared to non-holiday day.

**weekday**: We can see little difference in demand for shared bikes between days in week.

**workingday**: Booking trends seemed to be almost equal between working and non-working day.

**year**:2019 has more demand than 2018, so demand is increasing year by year.

---

**Question 2.** Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)
**Total Marks:** 2 marks (Do not edit)
**Answer:** <Your answer for Question 2 goes below this line> (Do not edit)

**drop_first=True** helps in reducing the extra column created during the dummy variable creation and hence avoid redundancy of any kind and can reduce the correlations created among dummy variables.

---

**Question 3.** Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)
**Total Marks:** 1 mark (Do not edit)
**Answer:** <Your answer for Question 3 goes below this line> (Do not edit)

According to the pair plot above, we can see that the **temp** and **atemp** show a strong correlation with the target variable, and we can see a linear regression pattern on the graph between temp and target variable, atemp and target variable.

---

**Question 4.** How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)

**Total Marks:** 3 marks (Do not edit)
**Answer:** <Your answer for Question 4 goes below this line> (Do not edit)

We can validate the assumptions of Linear Regression Model on the training set by:

- **Normal Distribution of Error terms:** The error terms should be normally distributed. we can check by plotting graph for error terms.

- **Multicollinearity:** There should not be no significant multicollinearity between variables (The independent variables should not show high correlations) this can be validated by checking VIF

- **Linearity**: The relationship between dependent variable and a feature variable must be linear.

- **Homoscedasticity:** The error should not be constant along the values of the dependent variables. We can check by drawing a scatterplot with the residuals, or using Breusch Pagan Test.

---

**Question 5.** Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
**Total Marks:** 2 marks (Do not edit)
**Answer:** <Your answer for Question 5 goes below this line> (Do not edit)

After our final model, we can see the top 3 predictor variables:

**Temperature** (temp) has a coefficient value of 0.4258, which indicates that it is a strong predictor variable influence the shared bike demand.

**Year** (yr) has a coefficient value of 0.2357, that means the demand for shared bike will increase as year increase

**weather** (weathersit_snow) has a coefficient value of -0.2434, which indicates that if it a snow day that will affect the booking of shared bikes.

---

## General Subjective Questions

**Question 6.** Explain the linear regression algorithm in detail. (Do not edit)
**Total Marks:** 4 marks (Do not edit)
**Answer:** Please write your answer below this line. (Do not edit)

<Your answer for Question 6 goes here>

**Linear regression** is a basic form of machine learning in which we train a model to predicting a dependent variable(target) based on the given independent variable(s). The regression technique tends to establish a linear relationship between a dependent variable and the other given independent variables. There are two types of linear regression- simple linear regression and multiple linear regression. Simple linear regression is used when a single independent variable is used to predict the value of the target variable. Multiple Linear Regression is when multiple

independent variables are used to predict the numerical value of the target variable. A linear line showing the relationship between the dependent and independent variables is called a regression line.

The linear regression model should provide a sloped straight line that represents the relationship between the variables.

Mathematically, the linear regression equation can be written as: $y=a+bx$

In which the formula of $a$ and $b$ can be:

$b=n\Sigma xy-(\Sigma x)\ (\Sigma y)/n\Sigma x2-(\Sigma x)2$

$a=n\Sigma y-b\Sigma x/n$

$y$ – Dependent variable (Target variable)

$x$ – Independent variable (Predictor variable)

$a$ – Interceptor of the line

$b$ – Slope of the line

The goal of the linear regression algorithm is to get the best value for $a$ and $b$, or known as finding the best fit line, the best fit line should have the least error.

---

**Question 7.** Explain the Anscombe's quartet in detail. (Do not edit)
**Total Marks:** 3 marks (Do not edit)
**Answer:** Please write your answer below this line. (Do not edit)

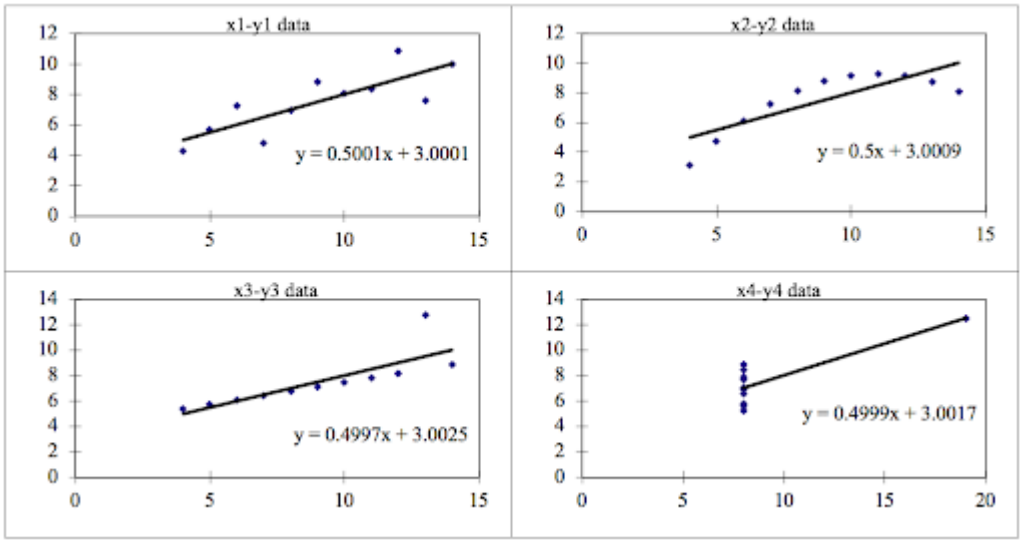
<Your answer for Question 7 goes here>

Anscombe's quartet consists of four data sets that have nearly identical simple descriptive statistics but have very different distributions and appear very different when presented graphically. Each dataset consists of eleven points. The primary purpose of Anscombe's quartet is to illustrate the importance of looking at a set of data graphically before beginning the analysis process as the statistics merely does not give an accurate representation of two datasets being compared.

Let's look at Anscombe's Data:

| Anscombe's Data | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Observation | x1 | y1 | | x2 | y2 | | x3 | y3 | | x4 | y4 |
| 1 | 10 | 8.04 | | 10 | 9.14 | | 10 | 7.46 | | 8 | 6.58 |
| 2 | 8 | 6.95 | | 8 | 8.14 | | 8 | 6.77 | | 8 | 5.76 |
| 3 | 13 | 7.58 | | 13 | 8.74 | | 13 | 12.74 | | 8 | 7.71 |
| 4 | 9 | 8.81 | | 9 | 8.77 | | 9 | 7.11 | | 8 | 8.84 |
| 5 | 11 | 8.33 | | 11 | 9.26 | | 11 | 7.81 | | 8 | 8.47 |
| 6 | 14 | 9.96 | | 14 | 8.1 | | 14 | 8.84 | | 8 | 7.04 |
| 7 | 6 | 7.24 | | 6 | 6.13 | | 6 | 6.08 | | 8 | 5.25 |
| 8 | 4 | 4.26 | | 4 | 3.1 | | 4 | 5.39 | | 19 | 12.5 |
| 9 | 12 | 10.84 | | 12 | 9.13 | | 12 | 8.15 | | 8 | 5.56 |
| 10 | 7 | 4.82 | | 7 | 7.26 | | 7 | 6.42 | | 8 | 7.91 |
| 11 | 5 | 5.68 | | 5 | 4.74 | | 5 | 5.73 | | 8 | 6.89 |
| | | | | Summary Statistics | | | | | | | |
| N | 11 | 11 | | 11 | 11 | | 11 | 11 | | 11 | 11 |
| mean | 9.00 | 7.50 | | 9.00 | 7.500909 | | 9.00 | 7.50 | | 9.00 | 7.50 |
| SD | 3.16 | 1.94 | | 3.16 | 1.94 | | 3.16 | 1.94 | | 3.16 | 1.94 |
| r | 0.82 | | | 0.82 | | | 0.82 | | | 0.82 | |

We can see 4 datasets share the same statistics summary (mean, standard deviation, r). However, after visualization, the data will look like below:




We can see that:
- 1st dataset (top left) appears to be a simple linear relationship between X and Y
- 2nd dataset (top right) shows a visible relationship between X and Y, but it is not linear.
- 3rd dataset (bottom left), the relationship is linear, but should have different line because of there is one outlier which can strongly affect the correlation coefficient.
- 4th dataset (bottom right) has a high leverage point that can produce a high correlation coefficient, even though the other data points do not demonstrate any relationship between the variables.

In conclusion, **Anscombe's quartet** helps us understand the important of data visualization before building a machine learning model, as it can easily fool a regression algorithm.

---

**Question 8.** What is Pearson's R? (Do not edit)
**Total Marks:** 3 marks (Do not edit)
**Answer:** Please write your answer below this line. (Do not edit)

<Your answer for Question 8 goes here>

In statistics, the **Pearson correlation coefficient**, also known as **Pearson's R**, the **Pearson product-moment correlation coefficient** (PPMCC), the **bivariate correlation**, or simply the **correlation coefficient**, is a measure of **linear correlation** between two sets of data. It is the most common way of measuring a linear correlation. It is a number between -1 and 1 indicates that it can measure the strength and the direction of the relationship between two variables.

| Pearson correlation coefficient ($r$) | Correlation type | Interpretation | Example |
|---|---|---|---|
| 0<r<1 | Positive correlation | When one variable changes, the other variable changes in the **same direction(both increase or decrease)** | Height and weight. The higher a person, the heavier his/her weight.. |
| 0 | No correlation | There is **no relationship** between the variables. | The price of a car is not related to the width of its windshield wipers. |
| -1<r<0 | Negative correlation | When one variable changes, the other variable changes in the **opposite direction**(one increases, the other decreases | Elevation & air pressure: |

---

**Question 9.** What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
**Total Marks:** 3 marks (Do not edit)
**Answer:** Please write your answer below this line. (Do not edit)

<Your answer for Question 9 goes here>

Feature scaling is a technique to standardize the predictor variables in a fixed range. It should be performed during the data pre-processing step to handle highly varying magnitudes or values or units. If it is not performed, machine learning algorithm tends to weigh greater values higher and consider small values as the lower values regardless of the unit of the values.

Example: In housing dataset, the number of bedrooms is ranged from 1 to 5, while the area feature has a wider range, can be up to 500 square meters. Machine learning algorithm can consider 500 to be greater than 1 or 5 but it is not true, so the prediction could be wrong.

Two most used techniques to perform Feature Scaling

**Min-Max Normalization**: Rescale a feature or observation value with distribution between 0 and 1

$$Xnew=Xi-min(X)/\max(X)-min(X)$$

**Standardization**: Rescale a feature value so that it has distribution with the mean of 0 and variance of 1

$$Xnew=Xi-X/\sigma x$$

Differences between Min-Max Normalization and Standardization:

| Min-Max Normalization | Standardization |
|---|---|
| Use minimum and maximum value of features for scaling | Use mean and standard deviation for scaling |
| Used when features are of different scales | Used when we want to ensure zero mean and unit standard deviation |
| Scale value between [0,1] or [−1,1] | Not bound to certain range |
| Affected by outliers | Much less affected by outliers |

---

**Question 10.** You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
**Total Marks:** 3 marks (Do not edit)
**Answer:** Please write your answer below this line. (Do not edit)

<Your answer for Question 10 goes here>

**VIF (Variant Inflation Factor)** helps explain the relationship of one independent variable with all the other independent variables. The formular of VIF is described as below:

1/(1-R2)

The value of **VIF is infinite** when there is a perfect correlation between the two independent variables. The R-squared value is 1 in this case. This leads to VIF infinity as VIF equals to 1/(1-R2). This concept suggests that is there is a problem of multi-collinearity and one of these variables need to be dropped in order to define a working model for regression.

---

**Question 11.** What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit)

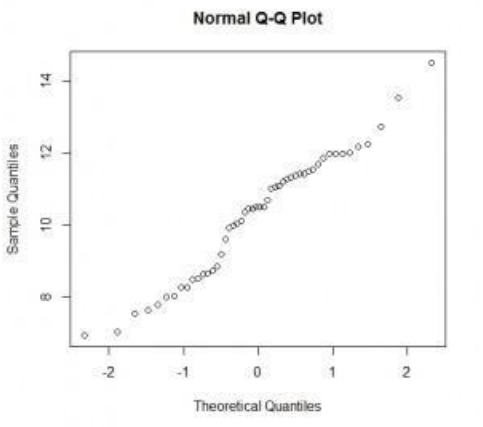
**Total Marks:** 3 marks (Do not edit)

**Answer:** Please write your answer below this line. (Do not edit)

<Your answer for Question 11 goes here>

The quantile-quantile (Q-Q) plot are used to plot quantiles of a sample distribution with a theoretical distribution to determine if any dataset concerned follows any distribution such as normal, uniform or exponential distribution. It helps us determine if two datasets follow the same kind of distribution. It also helps to find out if the errors in dataset are normal in nature or not.

This is an example of a Normal Q-Q plot when both sets of quantiles came from Normal distributions:




A 45-degree angle will be plotted on the Q-Q plot if the two datasets came from a common distribution, then the points will fall on that reference line.

In linear regression, Q-Q plot can help in the scenario when we have training and test dataset received separately and then we use Q-Q plot to ensure they are from populations with the same distribution. This can explain the importance of Q-Q plot in linear regression, it can detect many distributional aspects like shifts in location, shifts in scale, changes in symmetry, or the presence of outliers.

---