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)

   Answer:
   After fitting the LR model following categorical variables were affecting the count of bookings made

Categorical variables	**Effect on bike demand**
weathersit_Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist	cnt was lower (negative correlation)
weathersit_Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds	cnt was lower (negative correlation)
season_spring	cnt was lower (negative correlation)

   Count was lower for above categorical variables which means they affected the renting of bikes

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)
   Answer:
   Overview: Pandas function *get_dummies* is used to encode categorical variables to make it easier for ML algorithms to interpret relationships between categorical variables.

   drop_first=True - will drop the first category from encoding
   drop_first=False - will include all categories while encoding. Default is *False*

   Why?
   drop_first is set to True to avoid multicollinearity between the categorical variables

   Ex.
   season in bike assessment once converted to dummy variables converts to columns
   spring, summer, winter
   and ***fall*** was dropped

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
   Answer:
   Column ***temp*** i.e temperature shows the highest positive correlation

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
   Answer:
   Assumptions of Linear regression
   1. **Linearity:** The scatter plots of each independent variable against the dependent variable (cnt) show a linear relationship.
   2. **Independence:** The plot of residuals over the order of observations does not show any clear pattern. This indicates that the residuals are independent.
   3. **Homoscedasticity:** The scatter plot of residuals against the predicted values show no pattern (constant variance).
   4. **Normality:** The Q-Q (Quantile Quantile) plot show the residuals falling approximately along the 45-degree line, indicating that the residuals are normally distributed. The histogram show normal distribution.

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

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
   Answer:
   Overview: Linear regression algorithm is the process of identifying and quantifying a relationship between a dependent/ target variable and one or more independent/ predictor variables. Linear regression algorithms are used for prediction or forecasting of dependent variables using the relationship between dependent and independent variables.
   Ex. Predict price of houses given data for house prices in the past 5 years based on analyzing information from variables like locality, amenities, area, proximity from public transportation, year in which house was built and other such variables.

Goal: The goal is to find the best-fitting straight line (in the case of simple linear regression) or hyperplane (in the case of multiple linear regression) that minimizes the sum of the squared differences between the observed values and the values predicted by the model.

Types of Linear regression algorithm:

- Simple Linear Regression: The goal is to find relationship between a target variable and a predictor/ independent variable
- Multiple Linear Regression: The goal is to find relationship between a target variable and multiple predictor/ independent variable

Linear regression algorithm:

1. Simple Linear regression:

For a single predictor (independent variable), the model is expressed by the equation:

$$y = \beta_0 + \beta_1 x + \epsilon$$

Where:

β_0: intercept (value of y when x=0)

β_1: Slope (change in y for a 1 unit change in x)

y: dependent/ target variable

x: independent/ predictor variable

ϵ: error term (difference between observed and predicted value)

2. Multiple Linear regression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon$$

Where:

$x_1, x_2, \ldots, x_n$: Independent variables.

$\beta_1, \beta_2, \ldots, \beta_n$: Coefficients representing the effect of each predictor on y

β_0: intercept (value of y when x=0)

ϵ: error term (difference between observed and predicted value)

3. Assumptions: Simple and Multiple linear regression should follow 4 key assumptions - linearity, independence, homoscedasticity, normality

4. *Steps:*

    a. Reading and understanding the data

    b. EDA/ Visualizing the data

    c. Data conversion - normalization, standardization of data

    d. Splitting the data into training and test data sets

    e. Scaling the data

    f. Build a linear model: Goodness of fit: F-Statistic, p-value, t-value, standard error are used to determine a best fitting model

    g. Residual analysis: determine error distribution of the model

    h. Make prediction based on the model

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

Answer:

Overview: Anscombe's quartet are a group of 4 graphs that have similar simple descriptive statistics although appear different when graphs are plotted.

Key points:

A. Anscombe's quartet have similar simple descriptive statistics like mean, variance, correlation between x & y, and the best fit line is same

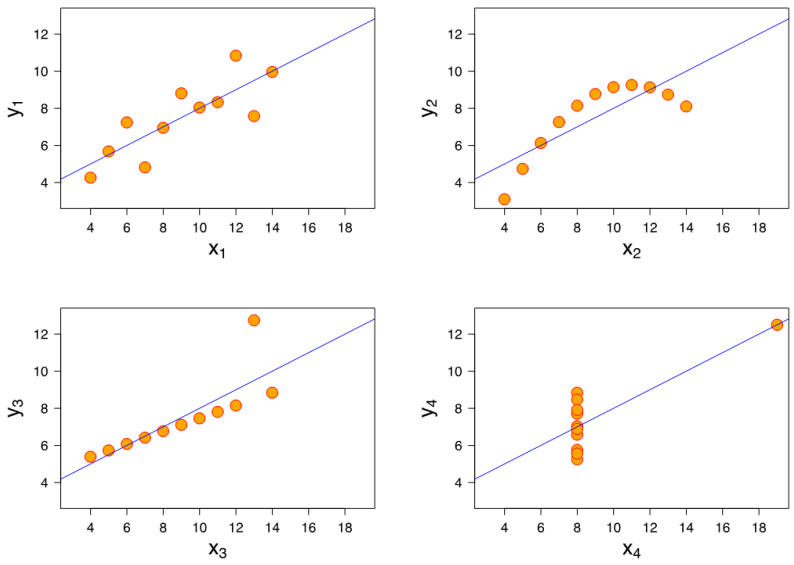
B. Anscombe's quartet when plotted shows different distributions of data and also brings forth problems like outliers and non-linearity

Ex. Following diagram shows 4 different graphs with the same best fit line but data points are distributed differently in all graphs.

A. *Graph 1:* shows linear relationship between x & y

B. *Graph 2:* Shows there is no linear relationship between x & y and other model can be applied to it

C. *Graph 3 & 4:* the best fit line would have been different had the outliers been removed but the presence of outliers changed the slope of the line altogether



Picture Credit: Wikipedia

Outcome: Anscombe's quartet shows the importance of
A. *Visual analysis*: Plot graphs wherever applicable to visualize the distribution of data w.r.t y
B. *Model Validation*: If the best fit line is correct in determining the model is a good fit or it is any other model should be applied
C. *Data validation*: Help to determine outliers and data point affecting the regression analysis

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

Answer:
Overview: Pearson's R is correlation coefficient to determine magnitude and direction of linear relationship between two continuous variables.
Magnitude lies between -1 & 1
Direction is either negative or positive
Formula:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Picture Credit: Google Search

Where:
x_i & y_i: are sample points
x_bar & y_bar: are mean of samples x & y respectively

Interpretation:
Correlation coefficient

1 -> *implies that x & y have perfect positive correlation i.e. change in variable x causes a positive proportional change in y*
-1 -> *implies that x & y have perfect negative correlation i.e. change in variable x causes a negative proportional change in y*

Application:
A. Pearson's R correlation coefficient is used in predictive analysis to determine magnitude and direction of correlation between variables
B. In Regression analysis it is predecessor to determine linearity of the model

Note: Pearson's R correlation coefficient does give an indication whether the relationship between variables is linear although Spearman's R is a better coefficient to determine the relationship is non-linear. It is possible to have a high Pearson's R coefficient value(ex. 0.7) but a higher Spearman's R coefficient(ex. 0.99) which concludes that the relation could be non-linear and a different model needs to be applied for this.

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

Answer:
Overview: Feature scaling is a method used to standardize or normalize the independent/ predictor/ feature variables so they are similarly distributed in a range.
Methods:
A. Standardization
B. Min Max Scaling (Normalization)
C. Log transformation

Reason(Why is scaling performed):
A. To increase the interpretability of the model. If the model adheres to one scale then comparison is easier
B. Scaled features/ variables results in faster convergence to minima in gradient descent algorithm
C. It also reduces computation time instead of calculating larger and distributed ranges, it is quicker to compute normally distributed data
D. Larger value ranges in features might create a bias in calculating the coefficients

Difference:

Standardization	**Normalization**

Scale feature variable to have mean 0 and standard deviation 1 (normally distributed data)	Scale feature variable to a fixed range typically (0,1)
$z = \frac{x - \mu}{\sigma}$ Where: x: original feature value μ: mean of feature values σ: standard deviation of feature values	$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$ Where: x: original feature variables x_min: minimum value of feature variables x_max: maximum value of feature variables
Outliers affect mean and standard deviation but the effect on scaling is lesser compared to Normalization	Outliers affect the min and max values of features and greatly affect the scaling of features
Standardization is used when data is assumed to be normally distributed and sensitive to feature scales	Normalization is used when data is not normally distributed or distribution of data is not known and features are on different scales

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

Answer:

Overview: VIF - Variance Inflation factor is used to determine whether multicollinearity between selected feature variables

Formula:

$$\text{VIF}(x_i) = \frac{1}{1 - R_i^2}$$

Where:
Ri2: R-Squared coefficient of determination of the regression of x_i on all the other predictors

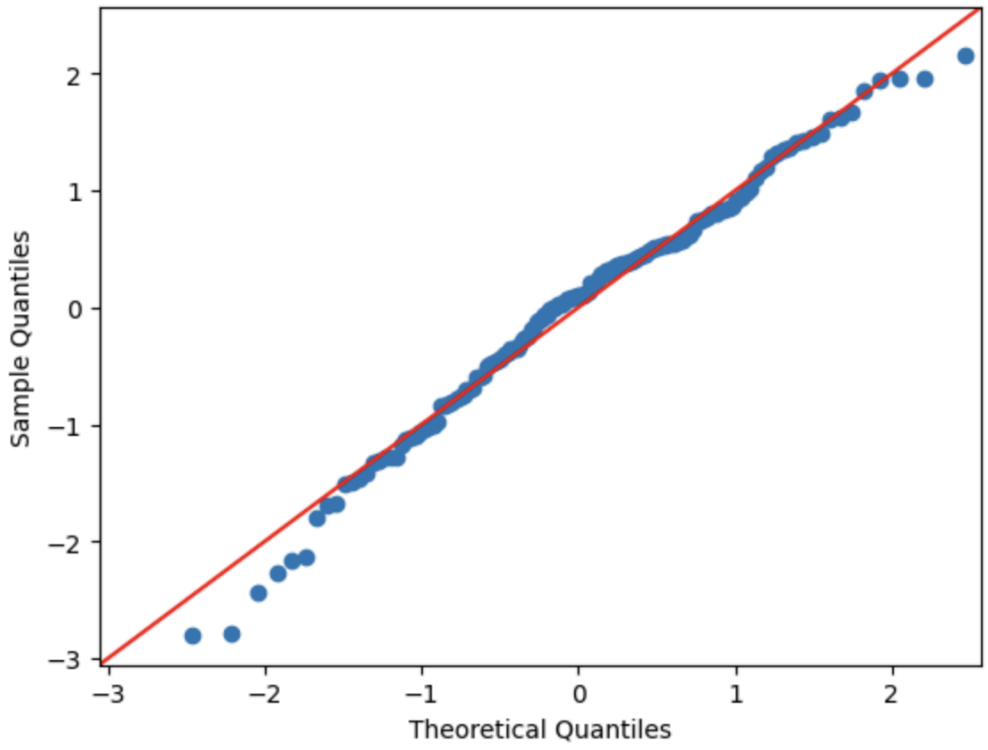
Interpretations:
A. VIF = 1 – No Multicollinearity
B. 1 < VIF < 5 – Moderate Multicollinearity
C. VIF > 5 – High Multicollinearity, Can be minimized
D. VIF > 10 – Very High Multicollinearity, potentially problematic model

Reason (Why VIF = ∞):

A. As per the formula VIF = ∞ – *When* Ri2 = 1 (R-squared value is 1)
B. *Interpretation:*
 i. Possibility of data containing duplicate feature variables or feature variables with nearly similar data points
 ii. One predictor being perfectly collinear to one or more predictor variables

How to fix:
A. Identify and remove the collinear feature variable
B. Combine variables to transform into a new feature variable

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
Answer:
Overview: QQ(Quantile - Quantile) plot - is used to determine if a dataset is normally distributed or not.



Interpretation of the Q-Q Plot

1. Diagonal Line: The red diagonal line represents the theoretical quantiles of the normal distribution.
2. Sample Points: The blue dots represent the sample quantiles from your dataset.

3. Normality Check: If the data is normally distributed, the points should lie approximately along the red line.