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?

ANS.

From analysis of the categorical variables from the dataset, following could be inferred about their effect on the dependent variable (count variable - DEMAND):

1. **DEMAND vs SEASON**

   For "fall" category median is the highest, followed by summer and winter. Spring season has the lowest median. The result is expected as temperature is optimal for fall and summer season for bike riding in America.

2. **DEMAND vs year (2018 and 2019)**

   Demand in year is 2019 is more as compared to that in 2018. Rent bikes might have become popular.

3. **DEMAND vs MONTHS (2018 & 2019)**

   July, Sep, June and Aug have high share of demand compared to other months as highlighted in season category that "fall" month has highest demand due to comfortable weather conditions.

4. **DEMAND vs HOLIDAYS**

   As expected, "no holidays" days have high demand than "holidays" indicating that rented bikes are popular among working people. However, "holiday" category has larger spread compared to non-holiday indicating higher preference for personal vehicles.

5. **DEMAND vs WEEKDAYS**

   Median is almost same for all the days with little difference with "Saturday" having more spread.

6. **DEMAND vs WORKING/NON-WORKING DAY**

   There is no difference in demand for both the categories.

7. **DEMAND vs WEATHER-SITUATION**

   Demand is high when weather is clear .

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

ANS.

drop_first=True is used to prevent multicollinearity. It helps in reducing extra column that is created during dummy variable creation.
For example: suppose we have 3 variables as “married”, “in relationship” and “unmarried”. We can take only 2 variables as – MARRIED = 1-0, IN – RELATIONSHIP – 0-1 and UNMARRIED can be dropped as 0-0, as it is understood.

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

ANS:

“temp” have the highest correlation with the target variable (count).

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

ANS:

Assumptions of Linear Regression are validated by:

1. NORMALITY OF ERROR TERMS : Plotting distplot of the residuals and analysing it to see if normal distribution is present or not and also check if mean is = 0.
2. MULTICOLLINEARITY : The assumption is that there is no significant multicollinearity among the features.
3. LINEARITY : This checks if the relationship between dependent and independent is linear or not. Residuals should be randomly scattered around zero when plotted against the predicted values.
4. Homoscedasticity or assumption that the variance of the residuals should remain constant across all the levels of the predictor variables.

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

ANS:
Following are the top 3 features that are highly correlated:

a. ‘temp’
b. ‘year’
c. ‘weather Light Snow & Rain’

This indicates that the bike rentals is majorly affected by temperature, season and month.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

ANS:

Linear regression is an algorithm that provides a linear relationship between an independent variable and a dependent variable to predict the outcome of future events. It is a statistical method used in data science and machine learning for predictive analysis.

It is a supervised machine learning method that finds a linear equation that best describes the correlation of the explanatory variables with the dependent variable.

This is achieved by fitting a line to the data using least squares. The line tries to minimize the sum of the squares of the residuals. The residual is the distance between the line and the actual value of the explanatory variable. Finding the line of best fit is an iterative process.

The following is an example of a resulting linear regression equation:

$$Y = b_0 + b_1x_2 + b_2x_2 + \ldots + b_nx_n$$

In the example above, y is the dependent variable, and x_1, x_2, and so on, are the explanatory variables. The coefficients (b_1, b_2, and so on) explain the correlation of the explanatory variables with the dependent variable. The sign of the coefficients (+/-) designates whether the variable is positively or negatively correlated. b_0 is the intercept that indicates the value of the dependent variable assuming all explanatory variables are 0.

Linear regression is a popular statistical tool used and has various benefits:

1. Easy implementation

The linear regression model is computationally simple to implement as it does not demand a lot of engineering overheads, neither before the model launch nor during its maintenance.

2. Interpretability

Unlike other deep learning models (neural networks), linear regression is relatively straightforward. As a result, this algorithm stands ahead of black-box models that fall short in justifying which input variable causes the output variable to change.

3. Scalability

Linear regression is not computationally heavy and, therefore, fits well in cases where scaling is essential. For example, the model can scale well regarding increased data volume (big data).

4. Optimal for online settings

The ease of computation of these algorithms allows them to be used in online settings. The model can be trained and retrained with each new example to generate predictions in real-time, unlike the neural networks or support vector machines that are computationally heavy and require plenty of computing resources and substantial waiting time to retrain on a new dataset. All these factors make such compute-intensive models expensive and unsuitable for real-time applications.

The above features highlight why linear regression is a popular model to solve real-life machine learning problems.

Q 2. Explain the Anscombe’s quartet in detail.

ANS:
Anscombe's quartet is a group of four datasets that have the same mean, standard deviation, and regression line, but which are qualitatively different.

They have different representations when we scatter plots on a graph.

It is often used to illustrate the importance of looking at a set of data graphically and not only relying on basic statistic properties.

The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.

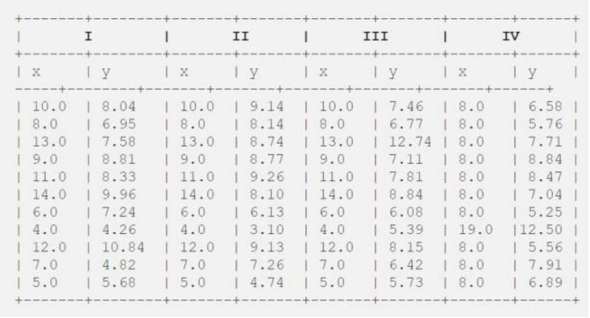
The four datasets that make up Anscombe’s quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.

Use of Anscombe’s quartet

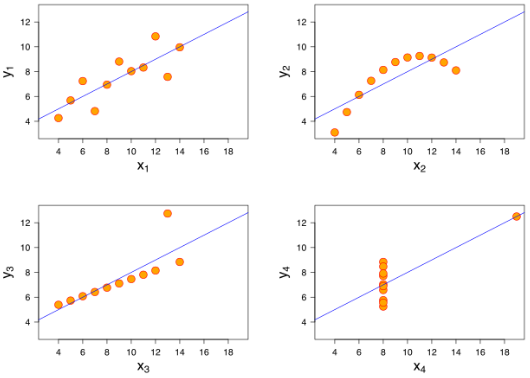
1. **It** is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics.
2. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.
3. It exhibits diverse patterns in scatter plots, illustrating the importance of visualizing data for meaningful insights beyond numerical summaries.
4. It Reveals limitations of summary statistics, emphasizing the need for visual exploration to detect nuances, outliers, and diverse relationships in datasets.
5. Anscombe’s Quartet underscores that numerical summaries alone can be misleading, emphasizing the crucial role of data visualization in uncovering patterns and outliers.

Anscombe’s Quartet Dataset

The four datasets of **Anscombe’s quartet.**



I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89



The four data sets **comprising the Anscombe's quartet:** all four sets have identical statistical parameters, but the graphs show them to be considerably different.

Q3. What is Pearson’s R?

ANS:

Pearson Correlation Coefficient or Pearson’s R

The Pearson correlation coefficient also known as Karl Pearson correlation coefficient, often denoted by the letter “r”, is a statistical measure that reflects the strength and direction of the linear relationship between two continuous variables.

It is **the most common way of measuring a linear correlation**.

Karl Pearson’s correlation coefficient formula is the most commonly used and the most popular formula to get the statistical correlation coefficient.

The formula for Pearson’s correlation coefficient is shown below:

$$r = n(\sum xy) – (\sum x)(\sum y) / \surd[n\sum x^2-(\sum x)^2][n\sum y^2-(\sum y)^2$$

Significance of Pearson’s R

The Pearson’s correlation helps in measuring the correlation strength (it’s given by coefficient r-value between -1 and +1) and the existence (given by p-value) of a linear correlation relationship between the two variables and if the outcome is significant we conclude that the correlation exists.

The Pearson correlation coefficient essentially captures how closely the data points tend to follow a straight line when plotted together. It’s important to remember that correlation doesn’t imply causation – just because two variables are related, it doesn’t mean one causes the change in the other.

According to Cohen (1988), absolute value of r of 0.5 is classified as large, an absolute value of 0.3 is classified as medium and an absolute value of 0.1 is classified as small.

Types of Pearson Correlation Coefficient

Each type of Pearson correlation coefficient offers unique insights and analytical tools for various research fields, from statistics and psychology to economics and engineering.

1. **Adjusted Correlation Coefficient**
   Adjusted correlation coefficient modifies the standard Pearson correlation coefficient to account for sample size and bias, especially when dealing with small sample sizes. It adjusts the correlation coefficient to provide a more accurate estimation of the population correlation.

2. **Weighted Correlation Coefficient**
   Weighted correlation coefficient assigns different weights to individual data points based on their importance or reliability. This approach is useful when certain observations carry more significance or have different levels of precision.
3. **Reflective Correlation Coefficient**
   Reflective correlation coefficient evaluates the relationship between variables in a reflective model, commonly used in structural equation modelling (SEM) to analyse latent constructs. It assesses the relationship between observed variables and underlying constructs.

4. **Scaled Correlation Coefficient**
   Scaled correlation coefficient scales the correlation coefficient to a specific range or magnitude, facilitating comparison across different datasets or studies. It ensures consistency in interpretation by standardizing correlation values.

5. **Pearson's Distance**
   Pearson's distance measures the dissimilarity or similarity between two data points based on their correlation coefficient. It quantifies the extent of deviation from perfect correlation, providing insights into the relationship between variables.

6. **Circular Correlation Coefficient**
   Circular correlation coefficient assesses the relationship between circular variables, such as angles or directions. It accounts for the cyclical nature of data and measures the degree of association between circular datasets.

7. **Partial Correlation**
   Partial correlation evaluates the relationship between two variables while controlling for the effects of one or more additional variables. It measures the unique association between variables after accounting for the influence of other factors, allowing researchers to isolate specific statistical relationships.

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

ANS :

SCALING

It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

SCALING IS PERFORMED –

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude. 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-squared**, etc.

Difference between normalized and standardized scaling

1. Normalised scaling brings all of the data in the range of 0 and **sklearn.preprocessing.MinMaxScaler** helps to implement normalization in python.

2. Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (**μ**) zero and standard deviation one (**σ**).

NORMALIZATION

1. This method scales the model using minimum and maximum values.
2. It is functional, when features are on various scales.
3. Values on the scale fall between [0, 1] and [-1, 1].
4. Additionally known as scaling normalization.
5. It is helpful when feature distribution is unclear.

Formula :

X_new = (X_min)/(x_max – X_min)

STANDARDIZATION

1. This method scales the model using the mean and standard deviation.
2. It is beneficial When a variable's mean and standard deviation are both set to 0.
3. Values on a scale are not con strained to a particular range.
4. This process is called Z-score normalization.
5. It is helpful, when feature distribution is consistent.

Formula:

X_new = (X – mean) / Std

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

ANS:

The value of VIF is calculated by the below formula:

$$VIF_i = 1 / (1 - R_i^2)$$

Where, ‘i’ refers to the ith variable.

If R-squared value is equal to 1 then the denominator of the above formula become 0 and the overall value become infinite. It denotes perfect correlation in variables.

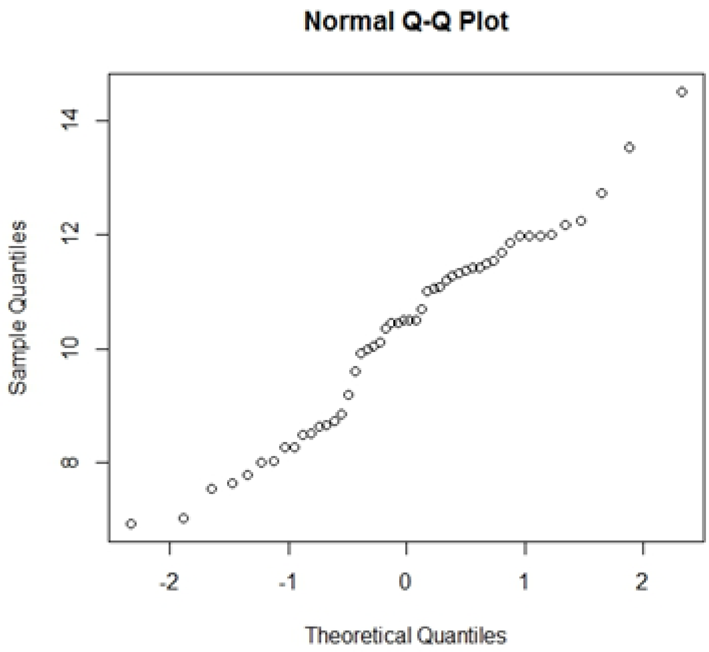
A variance inflation factor (VIF) is a measure of the amount of multicollinearity in regression analysis. Multicollinearity exists when there is a correlation between multiple independent variables in a multiple regression model. This can adversely affect the regression results. Thus, the variance inflation factor can estimate how much the variance of a regression coefficient is inflated due to multicollinearity.

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

ANS:

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

A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that’s roughly straight. Here’s an example of a Normal Q-Q plot when both sets of quantiles truly come from Normal distributions.



Use of Q-Q plot in Linear Regression: The Q-Q plot is used to see if the points lie approximately on the line. If they don’t, it means, our residuals aren’t Gaussian (Normal) and thus, our errors are also not Gaussian.

Importance of Q-Q plot: Below are the points:

I. The sample sizes do not need to be equal.

II. Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers.

III. The q-q plot can provide more insight into the nature of the difference than analytical methods.