Bike Sharing Assignment Assessment

by Irfan Khan Mohammed

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)**

- Insights shows the relationship between categorical variables and a Target variable on Bike Rentals are more:
    - In Category Year → The Bike Rental Business in 2019 had significantly increased compared to 2018 that is showing very positive business.
    - In Category Holiday → The Bike Rental Business on holiday had not been that effective as the variance is more random and scattered.
    - In Category Workingday → The Bike Rental Business on workingday was on higher side compared to non-workingday.
    - In Category Month → The Bike Rental Business seems high in September, followed by October, August, June, May, and April.
    - In Category Weathersit → The Bike Rental Business was on peak in Clear weather followed by Mist + Cloudy weather.
    - In Category Season → The Bike Rental Business was high in Fall Season followed by Summer Season.
    - In Category Weekday → The Bike Rental Business was high Saturday followed by Wednesday and Thursday.

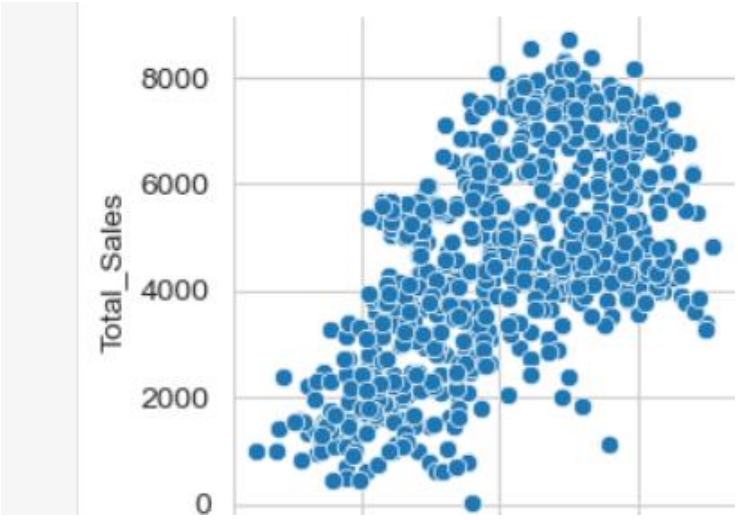
2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

- The purpose of drop_first=True is usually to avoid multicollinearity.
- In one-hot encoding, we set 1 to the position associated to a discrete value among n possible options.
- With one-hot encode a value, there is redundant information to figure out the value of any of the positions by computing 1 minus the sum of all other values.
- This means that any position of the one-hot encoded variable is a linear combination of the other positions.
- This linear correlation, however, can be a problem in some cases. One example is when we want to know the effect in the input features have on the prediction of a logistic or linear regression model.
- One solution to the multicollinearity of one-hot encoding is simply to remove one of the values.
- With that, we don't lose information, at the same time can remove the multicollinearity. drop_first=True is precisely for that.

---

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

- Temperature variable has highest correlation with the Total_Sales.



---

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

- The p-value for each term tests the null hypothesis that the coefficient is equal to zero (no effect) for all variables. A low p-value (< 0.05) indicates that you can reject the null hypothesis.

- A rule of thumb commonly used in practice is if a VIF is > 10, you have high multicollinearity. In our case, with values less than 5, we are in good shape, and can proceed with our regression.

- R-squared measures the strength of the relationship between your model and the dependent variable on a convenient 0 – 100% scale. And we have the R-square value of 0.835 or 83.5%.

- The adjusted R-squared adjusts for the number of terms in the model. And we got it around 0.832 or 83.2%.

- Our model is lr_model which is obtained by removing ("Mar", "June", "holiday", "Oct", "Wed", "Thu", "Aug", "Tue", "Mon", "May", "Feb", "humidity", "Sun", "Nov", "Dec", "Jan", "July", "Spring") variables using Statsmodel and VIF validations.

- Coefficient value of columns "Light Snow", "Mist + Cloudy" and "windspeed" is slightly negative, but P-value and VIF is within acceptable range.

- Finally, can conclude that from the above assumptions the VIFs and p-values both are within an acceptable range.

---

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)**

- Temperature could be a prime factor for making decision for the Organisation.
- We can see demand for bikes was more in 2019 than 2018.
- Working days as they have good influence on bike rentals. So, it would be great to provide offers to the working individuals.

General Subjective Questions

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

- Linear regression is a **type of supervised machine learning algorithm** that computes the **linear relationship between the dependent variable and one or more independent features** by fitting a linear equation to observed data.
- If there is **only one independent feature**, it is known as **Simple Linear Regression**, and if there are **more than one feature**, it is known as **Multiple Linear Regression**.
- Similarly, when there is **only one dependent variable**, it is considered **Univariate Linear Regression**, while when there are **more than one dependent variables**, it is known as **Multivariate Regression**.
- The interpretability of linear regression is a notable strength. The model's equation provides clear coefficients that elucidate the impact of each independent variable on the dependent variable, facilitating a deeper understanding of the underlying dynamics. Its simplicity is a virtue, **as linear regression is transparent, easy to implement, and serves as a foundational concept for more complex algorithms**.
- Linear regression is not merely a predictive tool; it forms the basis for various advanced models. Techniques like regularization and support vector machines draw inspiration from linear regression, expanding its utility. Additionally, linear regression is a cornerstone in assumption testing, enabling researchers to validate key assumptions about the data.

Types of Linear Regression: There are two main types of linear regression:

- **Simple Linear Regression**

  This is the simplest form of linear regression, and it involves only one independent variable and one dependent variable. The equation for simple linear regression is:
  $y=\beta0+\beta1X$ $y=\beta0+\beta1X$
  where:
  Y is the dependent variable
  X is the independent variable
  β0 is the intercept
  β1 is the slope

- **Multiple Linear Regression**

  This involves more than one independent variable and one dependent variable. The equation for multiple linear regression is:

$y=\beta 0+\beta 1X+\beta 2X+\ldots\ldots\beta nX$ $y=\theta 0+\theta 1X+\theta 2X+\ldots\ldots\theta nX$

where:

Y is the dependent variable

X1, X2, …, Xp are the independent variables

β0 is the intercept

β1, β2, …, βn are the slopes

The goal of the algorithm is to find the best Fit Line equation that can predict the values based on the independent variables.

---

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

- Anscombe's quartet comprises a set of **four datasets**, having identical descriptive statistical properties in terms **of means, variance, R-squared, correlations**, and linear regression lines but having different representations when we scatter plots on a graph.
- 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.
- **Purpose of Anscombe's Quartet**: Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. 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.

The four datasets of Anscombe's quartet sample:

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

---

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. Pearson's correlation (also called Pearson's R)** is a correlation coefficient commonly used in linear regression.
- 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. In simple terms, it answers the question, Can I draw a line graph to represent the data? Two letters are used to represent the Pearson correlation: Greek letter rho (ρ) for a population and the letter "r" for a sample.

Types of correlation coefficient formulas.

There are several types of correlation coefficient formulas.

One of the most commonly used formulas is Pearson's correlation coefficient formula. If you're taking a basic stats class, this is the one you'll probably use:

Pearson correlation coefficient

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\,n\Sigma x^2 - (\Sigma x)^2\,]\,[\,n\Sigma y^2 - (\Sigma y)^2\,]}}$$

Two other formulas are commonly used: the sample correlation coefficient and the population correlation coefficient.

Sample correlation coefficient

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

Sx and sy are the sample standard deviations, and sxy is the sample covariance.

Population correlation coefficient

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

The population correlation coefficient uses σx and σy as the population standard deviations, and σxy as the population covariance.

---

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

- Scaling 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 to bring all the variables to the same level of magnitude**. 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.

- 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.

The difference between normalized scaling and standardized scaling:

- **Normalization/Min-Max Scaling**: It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

  MinMax Scaling: x = x-min(x) \ max(x)-min(x)

- **Standardization Scaling**: 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). sklearn.preprocessing.scale helps to implement standardization in python.

  Standardization: x = x-mean(x) \ sd(x)

- One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

---

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

The value of VIF is calculated by the below formula:

$$VIF_i = \frac{1}{1-R_i^2}$$

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.

To interpret the VIF:

- The VIF is equal to 1 if the regressor is uncorrelated with the other regressors, and greater than 1 in case of non-zero correlation.
- The greater the VIF, the higher the degree of multicollinearity.
- In the limit, when multicollinearity is perfect (i.e., the regressor is equal to a linear combination of other regressors), the VIF tends to infinity.
- There is no precise rule for deciding when a VIF is too high (O'Brien 2007), but values above 10 are often considered a strong hint that trying to reduce the multicollinearity of the regression might be worthwhile.

---

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) plot is a graphical method for determining if a dataset follows a certain probability distribution or whether two samples of data came from the same population or not.** Q-Q plots are particularly useful for assessing whether a dataset is normally distributed or if it follows some other known distribution. They are commonly used in statistics, data analysis, and quality control to check assumptions and identify departures from expected distributions.
- Also, it helps to determine **if two data sets come from populations with a common distribution**.
- This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using **Q-Q plot that both the data sets are from populations with same distributions**.

- Few advantages of Q-Q plot
  a) It can be used with sample sizes also
  b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

- Quantiles are points in a **dataset that divide the data into intervals containing equal probabilities or proportions** of the total distribution. They are often used to describe the spread or distribution of a dataset. The most common quantiles are:

- **Quartiles (25th, 50th, and 75th percentiles): Quartiles divide the dataset into four equal parts**. The first quartile (Q1) is the value below which 25% of the data falls, the second quartile (Q2) is the median, and the third quartile (Q3) is the value below which 75% of the data falls.

Note:

- A **q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set**. For reference purposes, a 45% line is also plotted; For if the samples are from the same population, then the points are along this line.
