Bike Sharing Assignment

Submitted By : Snigdha Chakraborty

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?

6 categorical variables in the dataset was plotted using Box plot vs target variable and below points were inferred :

- **season**: Almost 32% of the bike booking is in season3 with a median of 5000 booking followed by season2 & season4 with 27% & 25% of total booking. This indicates, season can be a good predictor for the dependent variable
- **mnth**: Almost 10% of the bike booking is in the months 5,6,7,8 & 9 with a median of 4000 booking/month. This indicates, mnth has some trend with bookings and can be a good predictor for the dependent variable
- **weathersit**: Almost 67% of the bike booking is during weathersit1 with a median close to 5000 booking followed by weathersit2 with 30% of total booking. This indicates, weathersit has some trend towards the bike bookings can be a good predictor for the dependent variable
- **holiday**: Almost 97.6% of the bike booking is on non-holiday, the data seems clearly biased
- **weekday**: It shows close trend (between 13.5%-14.8% of total booking on all days of the week) with an independent medians between 4000 to 5000 bookings. This variable can have some or no influence towards the predictor, need to validate with further model building steps
- **workingday**: Almost 69% of the bike booking is on ‘workingday’ with a median of close to 5000 booking. This indicates, workingday can be a good predictor for the dependent variable

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

The key idea behind dummy encoding is that for a variable with, say, 'N' levels, we create 'N-1' new indicator variables for each of these levels. So for a variable say, 'Relationship' with three levels, namely, 'Single', 'In a Relationship', and 'Married' the dummy value will be

- Single – 1 0 0
- In A Relationship – 0 1 0
- Married – 0 0 1

We can see that we do not need 3 different levels and even if we drop first, we will still be able to explain the 3 levels

- Single – 0 0 (If both the dummy variables, namely, 'In a Relationship' and 'Married', are equal to zero, that means that the person is single)

To achieve the same in Python we use drop_first=True during dummy variable creation

Q2. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

The numerical variable which has the highest correlation with target variable is temp or atemp (temp and atemp are multicollinear and can be considered as same)

- temp : temperature in Celsius

- atemp: feeling temperature in Celsius

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

Following approach was taken to validate linear regression assumptions:

- Normality assumption: The distribution plot of error terms, ε(i), is normally distributed.
- Zero mean assumption: The residuals has a mean value of zero, i.e., the error terms are normally distributed around zero.
- Constant variance assumption: The residual terms have the same (but unknown) variance, σ2 . This assumption is also known as the assumption of homogeneity or homoscedasticity.
- Independent error assumption: The residual terms were independent of each other, i.e., their pair-wise covariance is zero.

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

Based on our final Model and all other inferences we concluded that the top 3 predictor variables that influences the bike booking are:

- Year, temperature and weather situation 3

And the next 2 important predictors are:

- season 4 and windspeed

Note :- The definition of variable as per data dictionary are as follows

- yr => Year (0: 2018, 1:2019)
- weathersit_3 => Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
- temp => Temperature in Celsius
- season_4 => Winter

General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

Linear regression is a machine learning algorithm based on **supervised learning** that finds the best linear-fit relationship on any given data, between independent and dependent variables. Regression models a target prediction value based on independent variables. It is mostly done by the Residual Sum of Squares Method.

Examples/Use Cases - Prediction of trends and Sales targets, price prediction etc.

If we take dependent variable as y and independent variable as x, then the equation of **best fit line** is presented as

$y = \beta 0 + \beta 1 * x$

While training the model we consider:

- x: input training data (univariate – one input variable(parameter))
- y: labels to data (supervised learning)
- β0: intercept
- β1: coefficient of x (slope)

Regression Hypothesis: Null hypothesis states that all slopes are equal to zero whereas alternate hypothesis states that at least one slope is not equal to zero. Mathematically it is represented as:

- **H0-> $\beta 1 = \beta 2 = \ldots = \beta n = 0$**
- **H1-> at least one $\beta i != 0$**

Q1. Explain the linear regression algorithm in detail (contd.)

Cost Function (J) : By achieving the best fit line the model aims to predict y value such that the error difference between actual and predicted value is minimum. To do that we need to update the coefficient or slope with every iteration till all the assumptions of linear regression is validated and the error between the predicted and actual value is minimized

The cost function of linear regression model is called as **Root Mean Squared Error (RMSE)** and mathematically represented as

$$J = \frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2$$

To minimize MSE/RMSE we use **Gradient Descent** to calculate the gradient of our cost function.

Assumption of Linear Regression:

It is assumed that there is a linear relationship between the dependent and independent variables. It is known as the ‘linearity assumption’.

Assumptions about the residuals: Normality assumption, Zero mean assumption, Constant variance assumption/Homoscedastic and Independent error assumption

Assumptions about the estimators: The independent variables are measured without error and The independent variables are linearly independent of each other, i.e., there is no multicollinearity in the data.

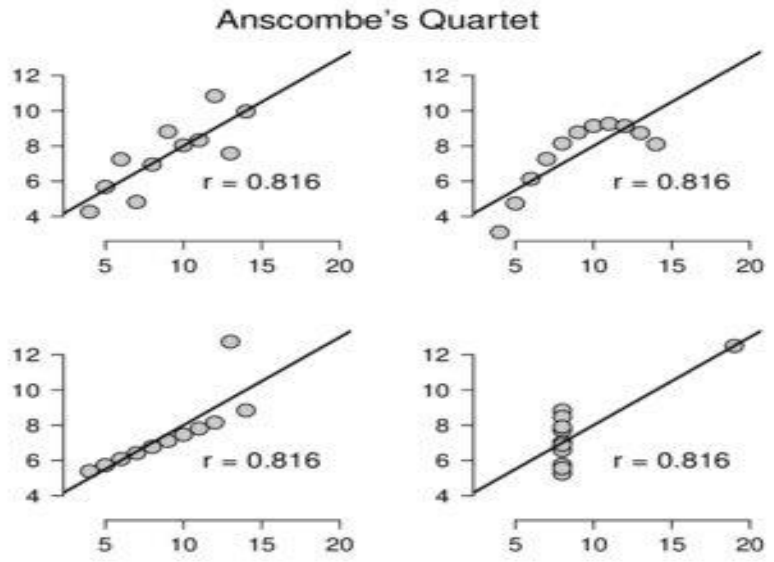
Q2. Explain the Anscombe’s quartet in detail.

Statistics are great for describing general trends and aspects of data, but statistics alone can't fully depict any data set. Francis Anscombe realized this in 1973 and created several data sets, all with several identical statistical properties. These data sets, collectively known as “Anscombe's Quartet,”

Anscombe's quartet comprises four data sets that have nearly identical statistics descriptions or summary but vary considerably when graphed using a distribution plot. They have a very different data distribution. Each data set consists of 11 (x,y) points. All four of these data sets have the same variance in x, variance in y, mean of x, mean of y, and linear regression.

It was constructed to demonstrate both the importance of graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties. His aim was to state that though the numerical calculations are exact, but the graphs are rough

The difference can easily be seen in the added image.



Summary

Set	mean(X)	sd(X)	mean(Y)	sd(Y)	cor(X,Y)
1	9	3.32	7.5	2.03	0.816
2	9	3.32	7.5	2.03	0.816
3	9	3.32	7.5	2.03	0.816
4	9	3.32	7.5	2.03	0.817

Q3. What is Pearson’s R?

Even if the two variables are correlated it does not signify that they have a linear relationship. It can be any relationship.

Correlation coefficient is majorly used in linear regression to validate whether there is a linear relationship between any of the independent variable with dependent variable. If not, the linear regression model will not yield good result.

There are 2 correlation coefficient that is used in regression

1. Pearson's R correlation coefficient
2. Spearman’s R correlation coefficient

The first one is used to depict linear relationship and second one is good for non-linear relationship

So, Pearson's R can be defined as a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

- Coefficient varies between -1 and +1 where:
  - r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
  - r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
  - r = 0 means there is no linear association
  - r > 0 < 5 means there is a weak association
  - r > 5 < 8 means there is a moderate association
  - r > 8 means there is a strong association

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

- Scaling is a method to have everything on the same scale for the model to be easily interpretable. It is performed during the data pre-processing to handle highly varying values or units. It normalizes the features of the data and is also known as data normalization.
- 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 resulting into incorrect modelling. To solve this issue, we must do scaling to bring all the variables to the same level of magnitude. In other words, algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values. Scaling only affects the coefficients and none of the other parameters such as F-Statistics, p-value, r-squared and so on.

Two ways of feature scaling are:

- **Normalized Scaling** – It is also known as min-max scaling. It brings all the data in the range of 0 to 1. sklearn.preprocessing.MinMaxScaler helps to implement the same in python.

  Min-Max Scaling(x) = (x – min(x)) / (max(x) – min(x))

- **Standardized Scaling** – It brings all 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.

  Standardized scaling(x) = (x – mean(x)) /std.dev(x)

The advantage of Standardization is that it doesn't compress the data between 0 to 1. This is useful, especially for outlier

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

- An infinite VIF(Variance Inflation Factor) value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.
- Mathematically :

  VIF = 1/ (1-R^2)

- In VIF, each feature is regression against all other features. If a feature is correlated with other features than R^2 will be more. If a feature is perfectly collinear, that is the behavior of one variable can depict the behavior of other the R^2 will tend to 1. Substitute R^ =1, VIF become

  VIF = 1 / (1 – R^2)

  = 1 / (1 – 1)

  = 1 / 0

  = infinity

- Hence, it’s concluded that when R^2 reaches 1, VIF reaches infinity.
- In short, If there is perfect correlation, then VIF is infinity.

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

- Quantile-Quantile (Q-Q) plot, is a plot of the quantiles of the first data set against the quantiles of the second data set. It is a graphical tool to help us assess if a set of data came from some theoretical distribution such as a Normal, exponential or Uniform distribution.
- When the training and test data set are received separately, and then Q-Q plot is used to confirm that both the data sets are from populations with same distributions. It is used to check following scenarios:
- If two data sets —
  - come from populations with a common distribution
  - have common location and scale
  - have similar distributional shapes
  - have similar tail behavior
- The assumption of normality is important in linear regression. The normal Q-Q plot is one way to assess normality. A 45 deg angle is plotted on the Q-Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.