Multivariate Regression:
Model Demand for Boom Bike Sharing

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?

- Season: There is higher demand for Boom bikes during summer and fall, with the lowest demand in spring.
- Year: The demand for boom bikes is trending higher year of year
- Month: Not surprisingly, the demand for boom bikes is highest in summer and fall months (from May through October)
- Holiday: There is more demand for Boom Bikes on regular days with greater dispersion is observed for holidays
- Weekday: No trend can be observed from the day of the week
- Weathersit: Better weather conditions, clear weather days have the highest demand for Boom Bikes

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

- Using **drop_first=True** removes the need for using an extra column for dummy variable creation. This reduces the redundancy and thus the correlation between the dummy variables. In some situations., it may be good to check if the dropped variable may could impact the interpretability of the model coefficients. One Hot encoding is a popular technique to convert categorical variables to numerical variables

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

- From the pair-plot, ‘Registered’ – that provides the count of the registered users has the highest correlation with the target variable – cnt. However, after feature selection, temp has the highest correlation as seen in the scatter plots when validating assumptions of linear regression

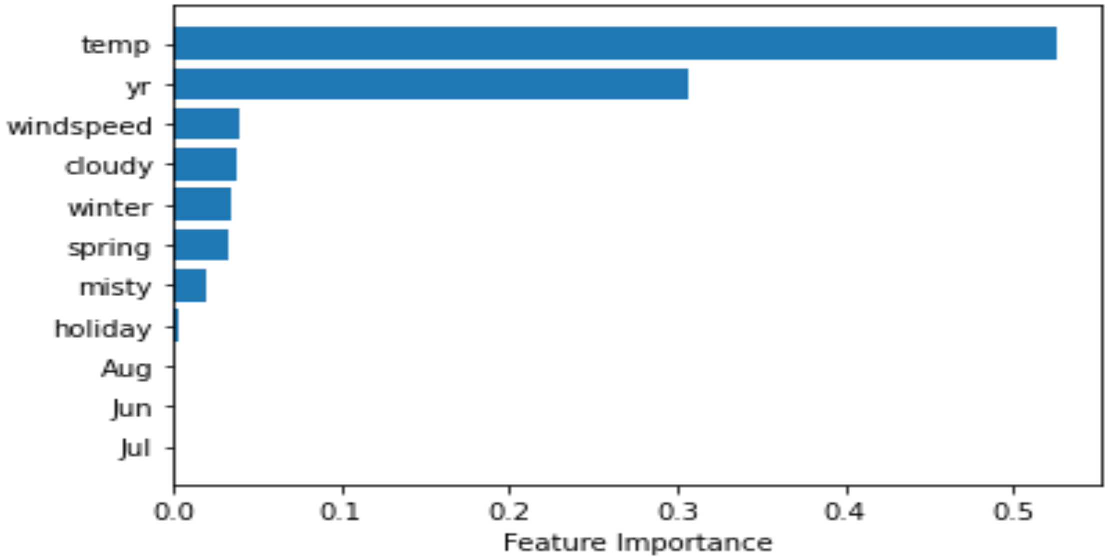
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

- Used scatter plots between features and the target variable to validate their linear relationship
- Checked for autocorrelation in residuals with Durbin Watson test
- Checked for heteroskedasticity by plotting residuals and fitted values

- Checked for multicollinearity- all features (other than the constant) had less than 5 in VIF values
- Checked for normality of residuals from the distribution plot of errors

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

- From the feature importance plot, we can see that ‘temp’, ‘yr’, followed by ‘windspeed’ are the top 3 features contributing to explain the demand of the shared bikes.



General Subjective Question

Q1. Explain the linear regression algorithm in detail

Linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables). The case of one explanatory variable is called *simple linear regression*; for more than one, the process is called multiple linear regression. This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable.
Ordinary least squares Linear Regression: Linear Regression fits a linear model with coefficients w = (w1, …, wp) to minimize the residual sum of squares between the observed targets in the dataset, and the targets predicted by the linear approximation.
class sklearn.linear_model.LinearRegression(***, *fit_intercept=True*, *normalize='deprecated'*, *copy_X=True*, *n_jobs=None*, *positive=False*)

Methods:

- fit(*X*, *y*, *sample_weight=None*): Fit linear model.
- predict(*X*): Predict using the linear model
- get_params(*deep=True*): Get parameters for this estimator
- score(*X*, *y*, *sample_weight=None*). Return the coefficient of determination of the prediction.

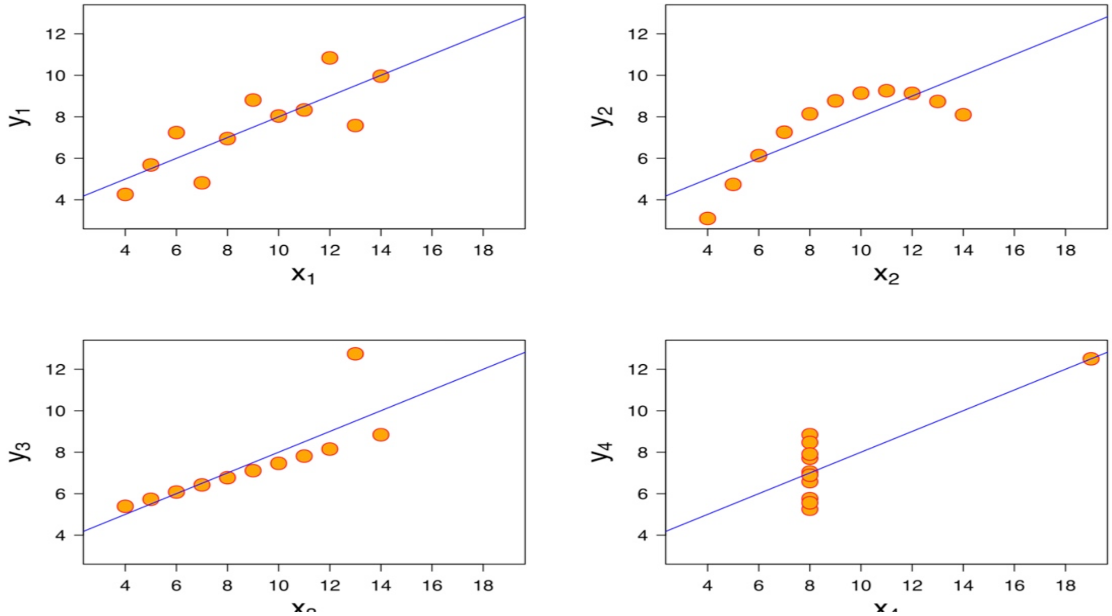
The coefficient of determination R2 is defined as (1−uv), where u is the residual sum of squares ((y_true - y_pred)** 2).sum() and v is the total sum of squares ((y_true - y_true.mean()) ** 2).sum(). The best possible score is 1.0 and it can be negative (because the model can be arbitrarily worse). A constant model that always predicts the expected value of y, disregarding the input features, would get a R2 score of 0.0

References:

- Wikipedia
- sklearn documentation on Linear. Regression

Q2. Explain the Anscombe's quartet in detail.

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (*x*,*y*) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analyzing it, and the effect of outliers and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough. This tells us about the importance of visualizing the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. Also, the Linear Regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets.



The four datasets can be described as:

1. **Dataset 1:** this **fits** the linear regression model well.

2. **Dataset 2:** this **could not fit** linear regression model on the data quite well as the data is non-linear.

3. **Dataset 3:** shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model

4. **Dataset 4:** shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model

References:

- Wikipedia
- https://towardsdatascience.com/importance-of-data-visualization-anscombes-quartet-way-a325148b9fd2

Q3. What is Pearson’s R

Pearson correlation coefficient — also known as Pearson's *r*, the Pearson product-moment correlation coefficient (PPMCC), the bivariate correlation, or colloquially simply as the correlation coefficient— is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus, it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1. As with covariance itself, the measure can only reflect a linear correlation of variables and ignores many other types of relationship or correlation. As a simple example, one would expect the age and height of a sample of teenagers from a high school to have a Pearson correlation coefficient significantly greater than 0, but less than 1 (as 1 would represent an unrealistically perfect correlation).

Interpretation:

The correlation coefficient ranges from −1 to 1. An absolute value of exactly 1 implies that a linear equation describes the relationship between *X* and *Y* perfectly, with all data points lying on a line. The correlation sign is determined by the regression slope: a value of +1 implies that all data points lie on a line for which *Y* increases as *X* increases, and vice versa for −1. A value of 0 implies that there is no linear dependency between the variables.

Python implementation:

DataFrame.corr(*method='pearson', min_periods=1*)
Compute pairwise correlation of columns, excluding NA/null values. Returns dataframe with pairwise correlation

References:

- Wikipedia
- Pandas documentation

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

Feature scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step. Since the range of values of raw data varies widely, in some machine learning algorithms, objective functions will not work properly without normalization. For example, many classifiers calculate the distance between two points by the Euclidean distance. If one of the features has a broad range of values, the distance will be governed by this feature. Therefore, the range of all features should be normalized so that each feature contributes approximately proportionately to the final distance. Another reason why feature scaling is applied is that gradient descent converges much faster with feature scaling than without it. It's also important to apply feature scaling if regularization is used as part of the loss function (so that coefficients are penalized appropriately)

Normalization:

Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.

Here’s the formula for normalization:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Here, Xmax and Xmin are the maximum and the minimum values of the feature respectively.

- When the value of X is the minimum value in the column, the numerator will be 0, and hence X’ is 0
- On the other hand, when the value of X is the maximum value in the column, the numerator is equal to the denominator and thus the value of X’ is 1
- If the value of X is between the minimum and the maximum value, then the value of X’ is between 0 and 1

Standardization:

Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

Here’s the formula for standardization:

$$X' = \frac{X - \mu}{\sigma}$$

$\mu$is the mean of the feature values and$\sigma$is the standard deviation of the feature values. Note that in this case, the values are not restricted to a particular range.

Normalization v/s Standardized Scaling

- Normalization is good to use when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbors and Neural Networks.
- Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.

References:

- Wikipedia
- https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/

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

Multicollinearity in regression analysis occurs when two or more explanatory variables are highly correlated to each other, such that they do not provide unique or independent information in the regression model. If the degree of correlation is high enough between variables, it can cause problems when fitting and interpreting the regression model.

Variance inflation factor (**VIF**) is the ratio (quotient) of the variance of estimating some parameter in a model that includes multiple other terms (parameters) by the variance of a model constructed using only one term. It quantifies the severity of multicollinearity in an ordinary least squares regression analysis. It provides an index that measures how much the variance (the square of the estimate's standard deviation) of an estimated regression coefficient is increased because of collinearity.

VIF is calculated as 1 / (1 – R Square)

An infinite VIF value indicates that a variable may be expressed exactly by a linear combination of other variables

If two independent variables have perfect correlation, then VIF = infinity. In such cases, R2 =1, thus 1/(1-RSquare) = infinity. To avoid this situation, we should remove one of the variables from the dataset that may be causing this perfect multicollinearity.

References:

- Wikipedia
- Statology

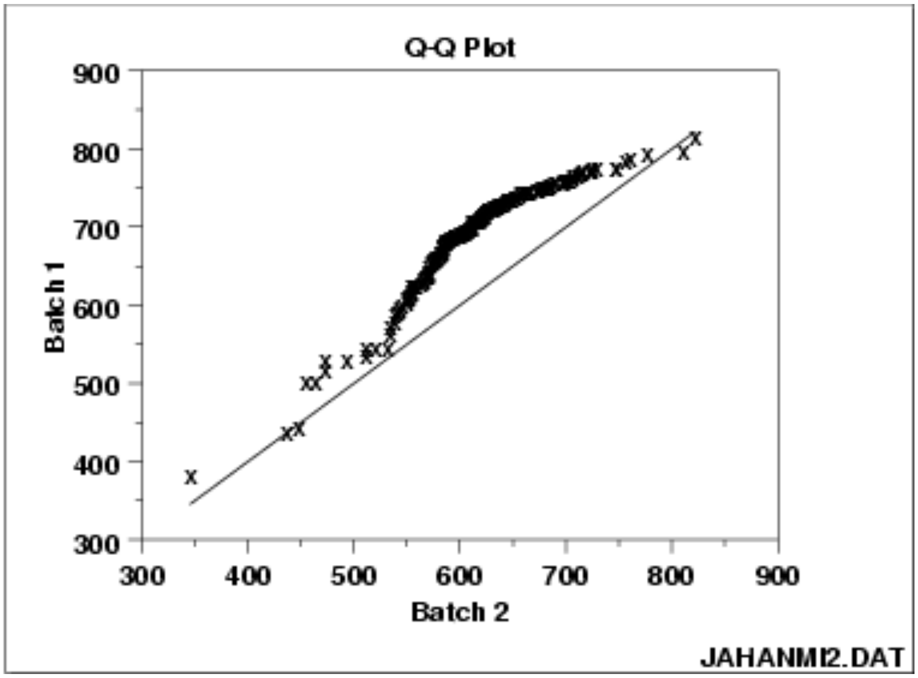
Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

The Q-Q or quantile – quantile plot is a graphical technique to determine if two datasets came from populations with a common distribution. Q–Q plot is a probability plot, used for comparing two probability distributions by plotting their quantiles against each other. First, the set of intervals for the quantiles is chosen. A point (*x*, *y*) on the plot corresponds to one of the quantiles of the second distribution (*y*-coordinate) plotted against the same quantile of the first distribution (*x*-coordinate). Thus, the line is a parametric curve with the parameter which is the number of the interval for the quantile.

If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line *y* = *x*. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line *y* = *x*.

Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions. Q–Q plots can be used to compare collections of data, or theoretical distributions. The use of Q–Q plots to compare two samples of data can be viewed as a non-parametric approach to comparing their underlying distributions. A Q–Q plot is generally a more powerful approach to do this than the common technique of comparing histograms of the two samples but requires more skill to interpret. Q–Q plots are commonly used to compare a data set to a theoretical model. This can provide an assessment of "goodness of fit" that is graphical, rather than reducing to a numerical summary. Q–Q plots are also used to compare two theoretical distributions to each other. Since Q–Q plots compare distributions, there is no need for the values to be observed as pairs, as in a scatter plot, or even for the numbers of values in the two groups being compared to be equal.



References:

- Wikipedia
- QQ Plot