Contents

Subjective Questions

Assignment-based Subjective Questions

From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Dependent variable or Target variable: cnt

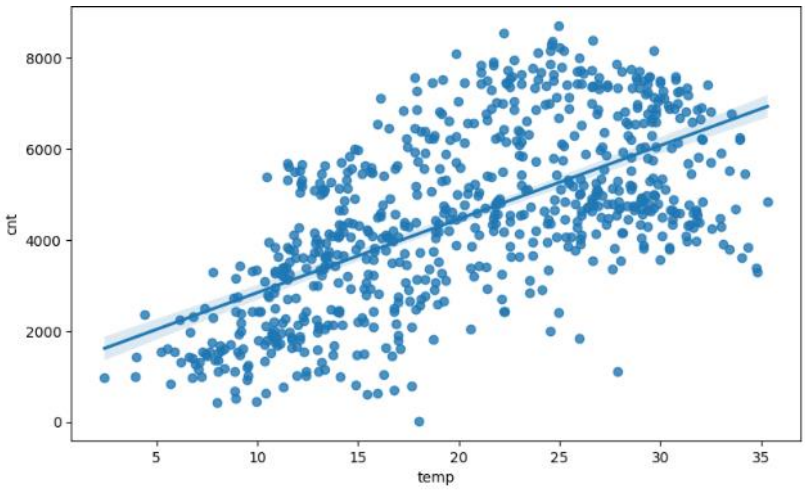
1. More for the Year 2019 compared to Year 2018. *{Based on box plot and bar plot}*
2. Most in clear/ less cloudy weather and least in light snow weather. *{Based on box plot and bar plot}*
3. With respect to seasons:
    - Much in the fall and Least in the Spring
    - Increases from spring to summer to fall and decreases from fall to winter. *{Based on box plot and bar plot}*
4. With respect to months:
    - The month September has highest, and Jan has lowest.
    - Increases from Jan to Jun, there is a dip in July then increase till September (peaked) decreases till the year end December. *{Based on box plot and bar plot}*
5. With respect to weekdays
    - Saturday, Wednesday, and Thursday bit more than other days however not much significant {Based on box plot – median}
    - There is an upward trend from Sun to Saturday (approximately) based on the bar plot visualization.
6. Average and Median of non-holidays rental is more than holidays *{Based on box plot and bar plot}*
7. Working day and non-working day shows almost same *{Based on box plot and bar plot}*
8. All variables exhibit upward trend from 2018 to 2019

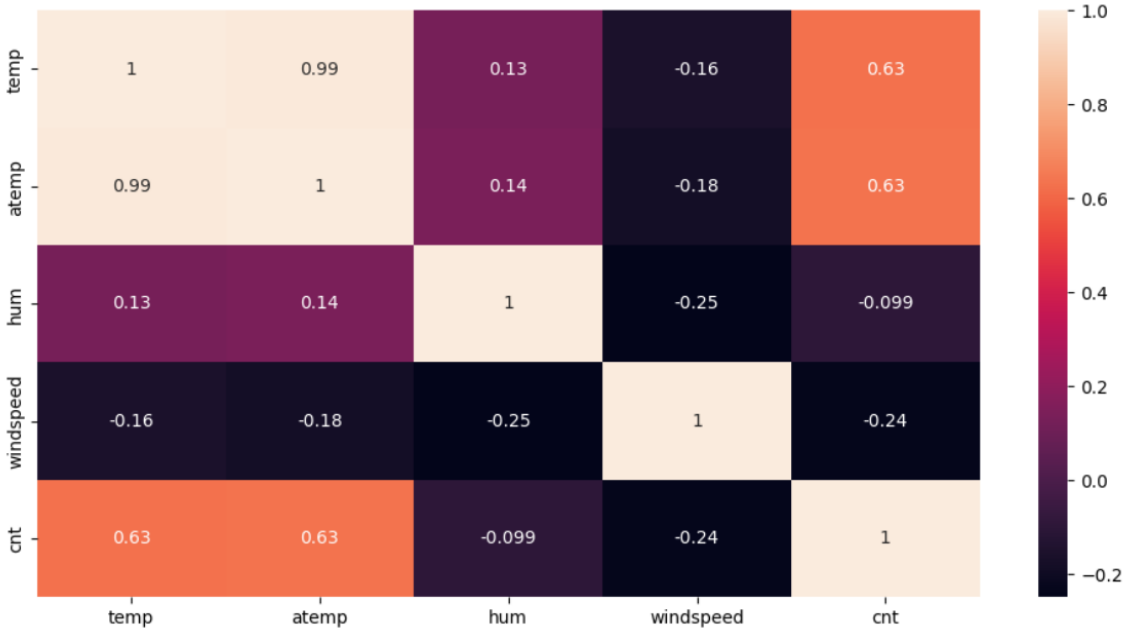
Why is it important to use drop_first=True during dummy variable creation?

- A dummy variable is a variable created to assign numerical value to levels of categorical variables. Each dummy variable represents one category of the explanatory variable and is coded with 1 if the case falls in that category and with 0 if not.
- Take our example Season, which has four categorical levels such as 1, 2, 3, 4 and as per the dictionary provided these are mapped {1:'spring', 2:'summer', 3:'fall', 4:'winter'}
- If a column has the value spring, we can encode the value as 1 for spring 0 for summer 0 for fall and 0 for winter. This can be done in python use get_dummies method from pandas' library.
- As we know if there are n variables and n-1 variables are zero then obviously the value belongs to nth variable. As per our example if the value is 0 for summer fall and winter, we can easily predict that it is spring. This leads to dependency between the columns. i.e Multicollinearity.
- To avoid the above scenario drop_first=True can be passed as one of the arguments in get_dummies(method).
- By doing this we drop the first categorical level hence we are reducing the number of columns and also the redundant information (multi collinearity).

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

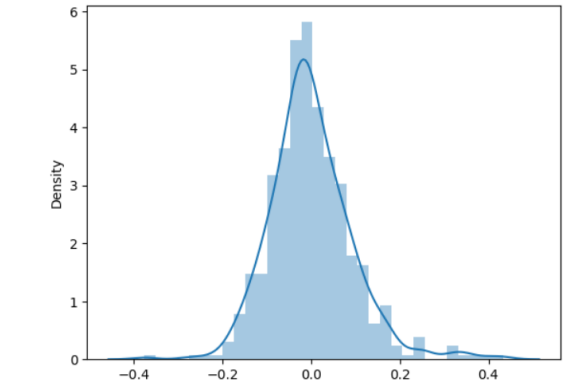
Temp indicator has the highest correlation. 0.63

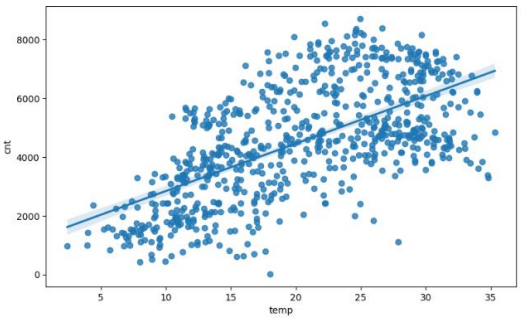


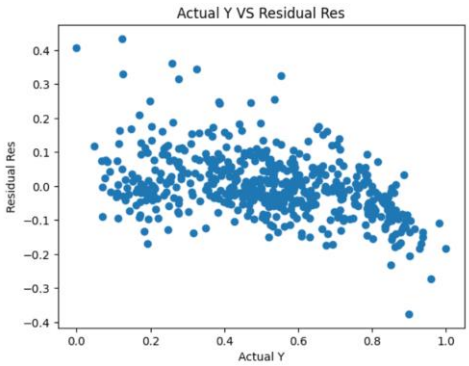


Note : Other indicators such atemp also has high correlation with cnt, but it has 99% i.e 0.99 correlation with temp, hence atemp is removed. Also registered, casual also have the correlation with cnt but those indicators were dropped in the model in the beginning itself (the 'cnt' variable indicates the total number of bike rentals, including both casual and registered)

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







- Target variable ‘cnt’ and prime impacting factor ‘temp’ are linear from the pair plot.
- Error terms are normally distributed (not X, Y), mean is zero (-4.4321844669349387e-16)
- Error terms are independent of each other i.e no pattern
- The indicators i.e independent variables are not correlated with each other, no multicollinearity among the independent variables. This is verified by checking the final VIF. All variables are less than 5
- Residual plots are done by keeping residues in Y-axis and target variable in X-Axis
- Error terms have constant variance (homoscedasticity)

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

Final Model

y = (0.0902) x *const* + **(0.0566)** x *workingday* + **(0.4914)** x *temp* + **(-0.0650)** x *season_spring* + **(0.0527)** x *season_summer* + **(0.0970)** x *season_winter* + **(0.0916)** x *mnth_Sep* + **(0.0645)** x *weekday_Sat* + **(-0.3041)** x *weathersit_LightSnow* + **(-0.0786)** x *weathersit_Mist* + **(0.2334)** x *yr_2019*

Coefficient table

	Coefficient
const	0.090189
workingday	0.056551
temp	0.491382
season_spring	-0.064952
season_summer	0.052659
season_winter	0.096999
mnth_Sep	0.091602
weekday_Sat	0.064533
weathersit_LightSnow	-0.304122
weathersit_Mist	-0.078644
yr_2019	0.233358

From the above model
Most positive impact i.e one unit increase in the indicator will increase the y value with their corresponding coefficient times.

- Temperature (0.4914)
- Year 2019 (0.2334)
- Winter Season (0.0970)

There are some negative impacting indicators also there such as

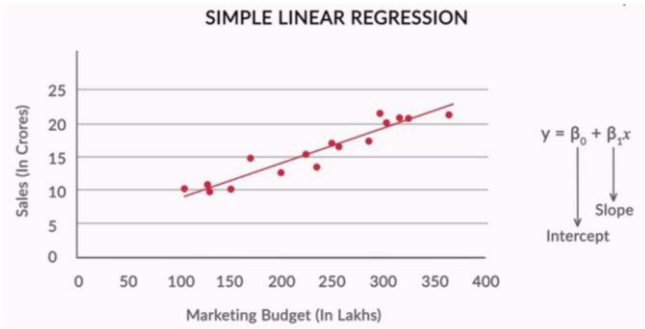
- LightSnow weather (-0.3041)

Negative impact i.e one unit increase in the indicator will decrease the y value their corresponding coefficient times.

General Subjective Questions

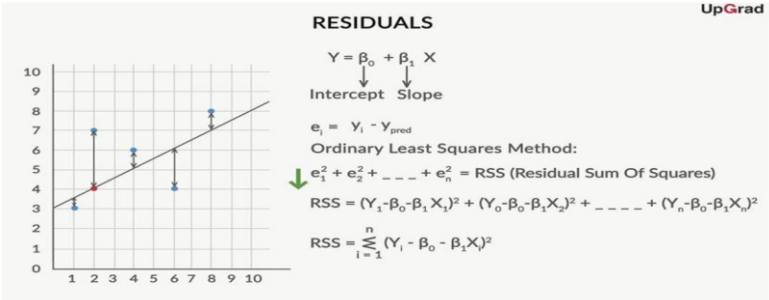
Explain the linear regression algorithm in detail.

- Regression is one of the three classified models in machine learning.
- The output variable to be predicted is a continuous variable.
- It is a supervised learning method.
- The most elementary type of regression model is the simple linear regression which explains the relationship between a dependent variable and one independent variable using a straight line.



X – axis represents dependent variable and Y- axis represents Dependent variable

- Residuals for any data point is found by subtracting predicted value of dependent variable from actual value of dependent variable.
- RSS (Residual Sum of Squares) is equal to the sum of squares of the residual for each data point in the plot.
- The best-fit line is found by minimising the expression of RSS (Residual Sum of Squares) which is equal to the sum of squares of the residual for each data point in the plot.



- The strength of the linear regression model can be assessed using 2 metrics.
    - R^2 or Coefficient of Determination
    - Residual Standard Error (RSE)

R^2 or Coefficient of Determination

$$R^2 = 1 - (RSS / TSS)$$

- RSS (Residual Sum of Squares): The residual sum of squares (RSS) measures the level of variance in the error term, or residuals, of a regression model.
  $RSS = \sum^{n}_{i=1} (y^{i} - f(x_i))^2$

- TSS (Total sum of squares): It is the sum of errors of the data points from mean of response variable. Mathematically, TSS is:

  $TSS = \Sigma(Yi - \text{mean of } Y)^2$

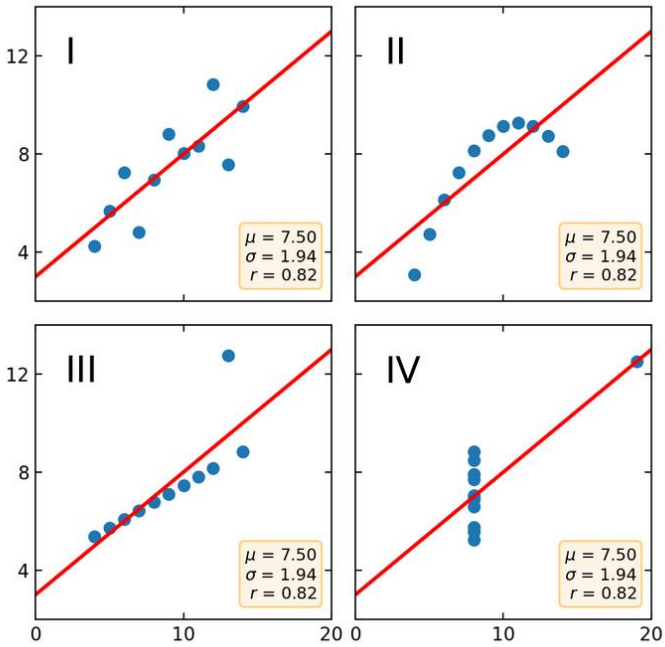
- R^2 - Higher the value the better the model fits your data.
- Assumptions of simple linear regression (SLR)
  - Linear relationship between X and Y
  - Error terms are normally distributed (not X, Y)
  - Error terms are independent of each other.
  - Error terms have constant variance (homoscedasticity
- One way to increase the value of R^2 adding more parameters to our model if available.
- Adding more indicators/ parameters/ features to the model and analysing is called Multiple linear regression (MLR).
- MLR is a statistical technique to understand the relationship between one dependent variable and several independent variables.
- However, sometimes higher R^2 value leads to a phenomenon called "Over fitting".
- "Over fitting" represents it is a very good fit for the trained data, but it may not be good for predicting the future values.
- "Multicollinearity" is another phenomenon we need to avoid, which is association between the predictor variables.
- Assumptions of multiple linear regression have similar to above mentioned for SLR
- Variance Inflation Factor (VIF) helps us to eliminate the inter dependencies among the variables. If VIF is greater than 5 that indicator can be eliminated because it has high correlation with other indicator(s).
- Feature selection is important for modelling, Recursive Feature Elimination (RFE) is one of the techniques help us to eliminate the not required features for the model. This is an automatic approach.
- After that we can perform manual approach to eliminate the features which has high p-value then high VIF. This process is to be repeated till we get lower or zero p-value or lower VIF value. Lower p-value means the variable is significant to the model.

Explain the Anscombe’s quartet in detail.

Anscombe's quartet is a group of datasets (x, y) that have the same mean, standard deviation, and regression line, but which are qualitatively different.
It conveys two important ideas.

- Dependence on statistics summary is not sufficient.
- Visualization is a key process for understanding the data qualitatively.

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.



Picture Ref : https://matplotlib.org/stable/gallery/specialty_plots/anscombe.html

In the above 4 data, we can see different distributions while plotting the graphs, but the summary statistics are small among the 4 datasets.

What is Pearson's R?

When there exists some relationship between two measurable variables, we compute the degree of relationship using the correlation coefficient.

Co-variance

Let (X,Y) be a bivariable normal random variable where V(X) and V(Y) exists. Then, covariance between X and Y is defined as

$$\text{cov}(X,Y) = E[(X-E(X))(Y-E(Y))] = E(XY) - E(X)E(Y)$$

If (xi, ,yi), i=1,2, ..., n is a set of n realisations of (X,Y), then the sample covariance between X and Y can be calculated from

$$\text{cov}\left(X,Y\right)=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})=\frac{1}{n}\sum_{i=1}^{n}x_i y_i-\bar{x}\,\bar{y}$$

When X and Y are linearly related and (X,Y) has a bivariate normal distribution, the co-efficient of correlation between X and Y is defined as

$$r\left(X,Y\right)=\frac{\text{cov}(X,Y)}{\sqrt{V(X)V(Y)}}$$

Pearson's correlation coefficient, often denoted as "r" or "Pearson's r,"
This is also called as product moment correlation co-efficient which was defined by Karl Pearson.
Based on a given set of n paired observations (xi,yi), i=1,2, ... n the sample correlation co-efficient between X and Y can be calculated from

$$r\left(X,Y\right)=\frac{\frac{1}{n}\sum_{i=1}^{n}x_i y_i-\bar{x}\,\bar{y}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}x_i^2-\bar{x}^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}y_i^2-\bar{y}^2}}$$

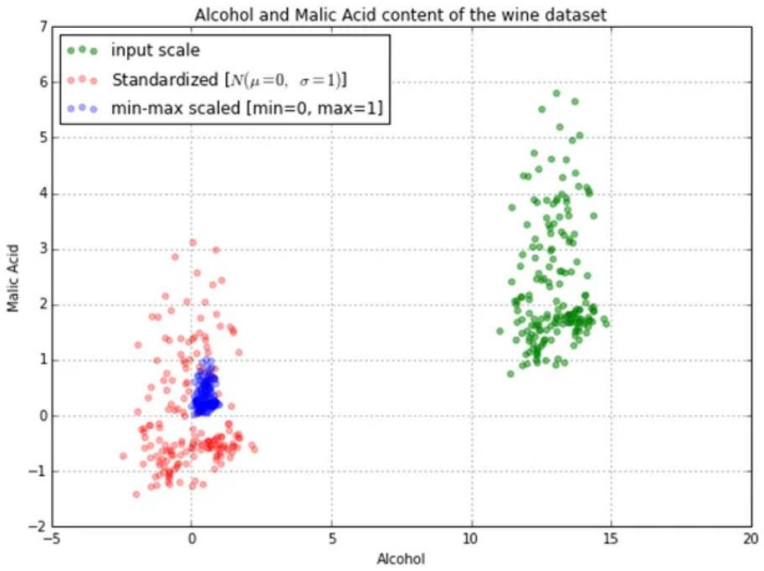
Notes on correlation coefficient

- The correlation coefficient between X and Y is same as the correlation coefficient between Y and X
- The correlation coefficient is free from the units of measurements of X and Y
- The correlation coefficient is unaffected by change of scale and origin.
- The correlation coefficient lies between -1 and +1.
- A positive value of 'r' indicates positive correlation.
- A negative value of 'r' indicates negative correlation.
- If r = +1, then the correlation is perfect positive
- If r = –1, then the correlation is perfect negative.
- If r ≥0.7 then the correlation will be of higher degree. In interpretation we use the adjective 'highly'

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

Feature Scaling:

- Feature scaling is the process of normalizing the range of features in a dataset, sometimes we have a lot of independent variables in a model, a lot of them might be on very different scales. Real-world datasets often contain features that are varying in degrees of magnitude, range, and units. Therefore, for machine learning models to interpret these features on the same scale, we need to perform feature scaling.
- Feature scaling is mainly performed for two reasons.
    - Ease of interpretation
    - Faster convergence for gradient descent methods
- There are two popular methods for scaling.
    - Standardizing:
        - The variables are scaled in such a way that their mean is zero and standard deviation is one.
    - MinMax Scaling:
        - The variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data.
- The advantage of Standardisation over the other is that it doesn't compress the data between a particular range as in Min-Max scaling. This is useful, especially if there are extreme data point (outlier).
- 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 impact of Standardization and Normalisation on the Wine dataset

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

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

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

- For a perfect correlation scenario, R^2 will become 1 and 1/ (1-1) = 1/0 will become infinite.
- Multicollinearity arises when a regressor is very similar to a linear combination of other regressors.
- The regressor is equal to a linear combination of other regressors, the VIF tends to infinity. This is called perfect multicollinearity.
- When the VIF is high, which means that predictor/ indicator itself can be predicted by other set of variables, so this variable should be dropped for further analysis.
- It is always good to analyse why is this dependency.

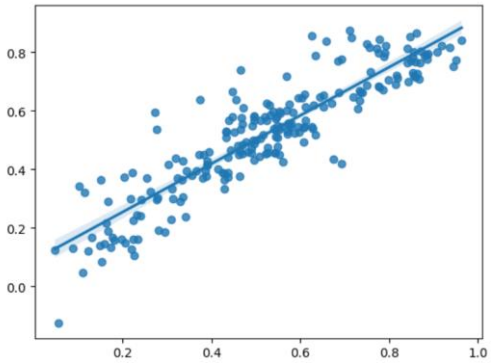
What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

- The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.
- A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

The advantages of the q-q plot are:

- The sample sizes do not need to be equal.
- Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot. For example, if the two data sets come from populations whose distributions differ only by a shift in location, the points should lie along a straight line that is displaced either up or down from the 45-degree reference line.

In our case, to check whether the test data and trained data have come from the same distribution we can use Q-Q plot



The q-q plot is formed by:

- Vertical axis: Estimated quantiles from data set 1 (Train set in our example)
- Horizontal axis: Estimated quantiles from data set 2 (Test set in our example)

References

- https://cdn.upgrad.com/UpGrad/temp/3fa7448c-58fd-4774-8192-a3d5e289b78e/Linear+Regression+Lecture+Notes.pdf
- https://towardsdatascience.com/significance-of-q-q-plots-6f0c6e31c626
- https://www.sigmamagic.com/blogs/what-is-variance-inflation-factor/#:~:text=If%20there%20is%20perfect%20correlation,to%20the%20presence%20of%20multicollinearity.
- https://drive.google.com/file/d/1TM9CK4x6w0XURKDNY2vLvAnXxhZqzGWc/view
- https://matplotlib.org/stable/gallery/specialty_plots/anscombe.html