Linear Regression Assignment

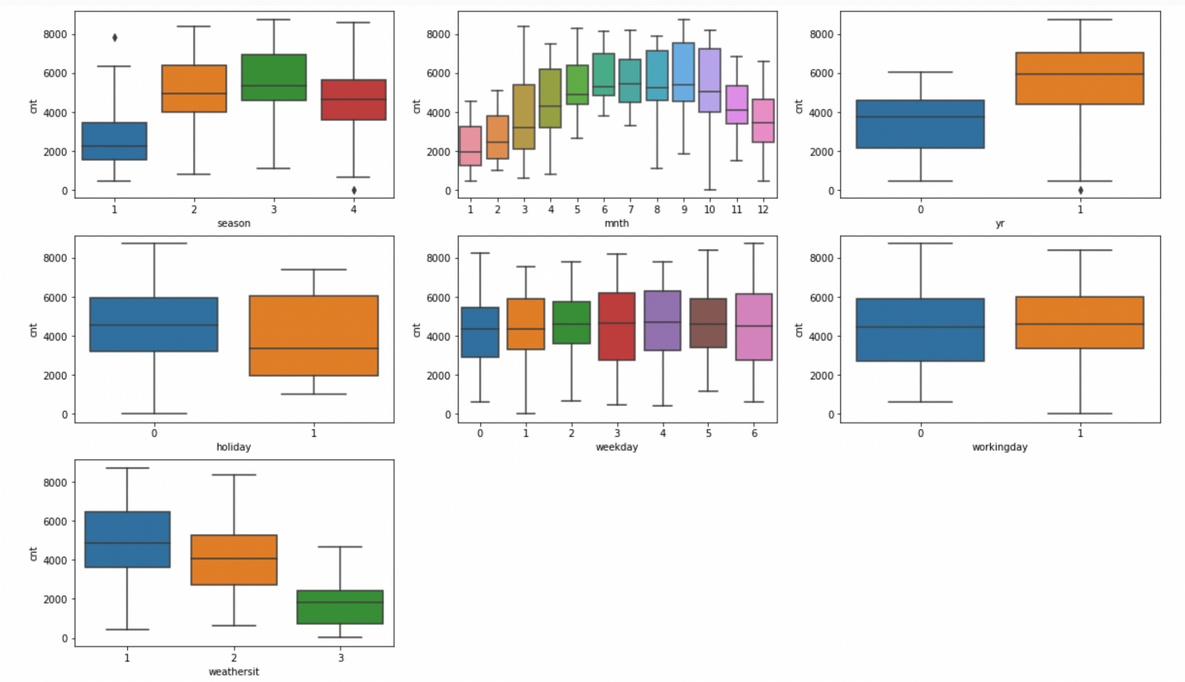
Gokul GS 20 Apr 2022

Subjective Questions

1. Effect of categorical variables on independent variable

List of Categorical variables

- Weekday - Has no impact on bike rental bookings
- Holiday - Mild impact, although we see from the box plot that more bookings are taken during non-holidays than holidays. Regression coeff -0.0860
- Season - Bookings increases from Seasons 1 to 3. Fall has maximum bookings follow by Summer and winter and spring
- Year - Significant jump in bookings in the year 2019 compared to 2018. Overall bike sharing is gaining popularity. Regression coeff 0.2307. Increase in bookings by 0.2307 units in 2019 vs 2018)
- Month - Bookings increase from Jan till May and peak during the months 6 to 9 before again dropping slowly till Dec
- Weather type - Maximum bookings happen in clear weather conditions and drop as the weather worsens



2. Using drop_first=True

drop_first=True ensures that there are only N-1 variables selected for any categorical variable with N values. This helps in reducing the total parameters used for analysis and simplifies understanding.
Eg: Weathersit has 4 values from 1.Clear sky, 2.Mist, 3.Light rain / snow and 4. heavy rain / snow.

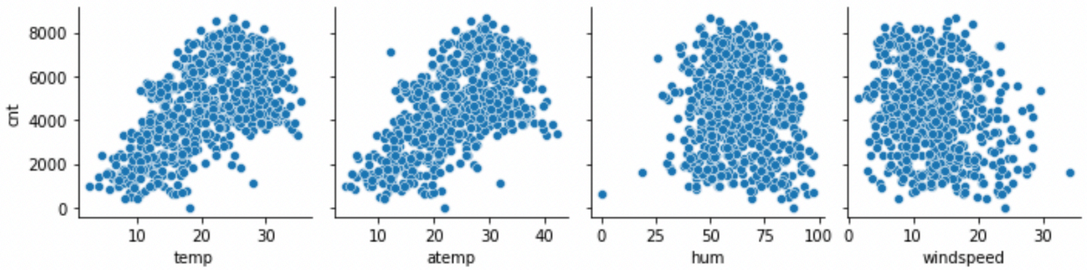
All 4 types can be identified using only 3 variables with binary values as shown below

	1 Clear sky	2 Mist	3 Light rain or snow
Clear Sky	1	0	0
Mist	0	1	0
Light Rain	0	0	1
Heavy Rain / Snow	0	0	0

3. Correlation with Target variable

- We find that temperature (“temp”) has the highest correlation with the Target variable “cnt”, bookings

```
In [294]: sns.pairplot(data=bookings,x_vars=["temp","atemp","hum","windspeed"],y_vars="cnt")
Out[294]: <seaborn.axisgrid.PairGrid at 0x7fd8a9031520>
```

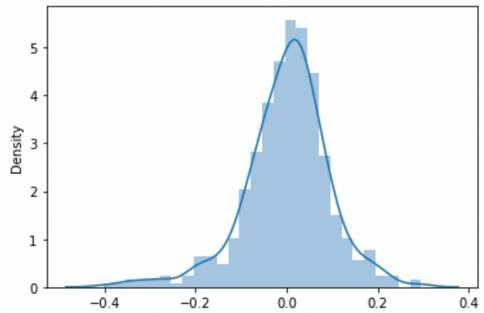


- We also observe that the variable “atemp” seems to be a derived metric of temp with a high correlation of 0.99. Hence consider only the variable “temp” in the model and drop “temp”

4. Validating the assumptions of linear regression

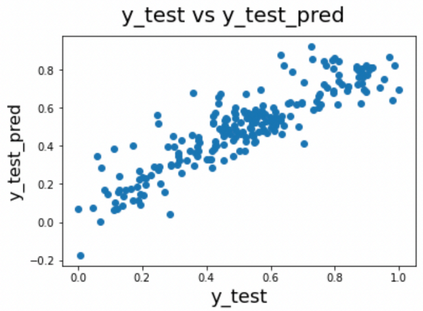
- Error terms are normally distributed

```
In [254]: #plot the residuals

          sns.distplot(residuals)

Out[254]: <AxesSubplot:ylabel='Density'>
```



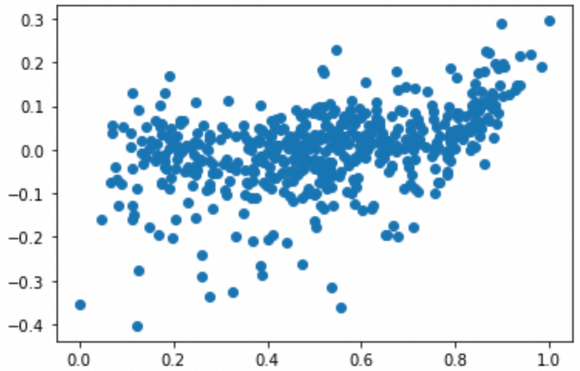
- X and Y are independent

```
In [278]: #understanding the spread
          fig = plt.figure()
          plt.scatter(y_test,y_test_pred)
          fig.suptitle('y_test vs y_test_pred', fontsize=20)
          plt.xlabel('y_test', fontsize=18)
          plt.ylabel('y_test_pred', fontsize=16)

Out[278]: Text(0, 0.5, 'y_test_pred')
```



- Error terms are independent - no pattern

```
In [296]: #check if residuals are independent - no pattern
          plt.scatter(y_train,residuals)
          plt.show()
```



- Homoscedasticity - Error terms have a constant variance - we can see that the variance remains constant except for higher values

5. Top 3 features explaining the demand for shared bikes

Equation

total_bookings = 0.2183+0.2307(year)*
+0.4960(temperature)-0.0860*(holiday)-0.1406*(humidity)-0.1830*(windspeed)*
+0.1180(summer_flag)+0.0749*(fall_flag)*
+0.1620(winter_flag)-0.0522*(mist_flag)-0.2396*(light snow or light_rain)*

The top 3 features that impact bike rental bookings are

1. Temperature (change of 1 unit increases bookings by 0.4960 units)
2. Light Snow or Light Rain (Weathersit=3) (if there is light snow the bookings drop by 0.2396 units)
3. Year (increase in bookings by 0.2307 units in 2019 vs 2018)

General Subjective Questions

1. Linear Regression algorithm

- It is a supervised ML algorithm used to understand a given data set and build predictors for forecasting.

Pre requisites :

- The algorithm needs prepared data, variables which are redundant are removed and dummy variables are created for categorical variables
- Once the data set is split into train and test set, the variables are transformed (scaling) to ensure uniformity and interpretability
- Use either the stats model approach or the recursive feature elimination approach. In both the cases the algorithm takes the least squares (cost function) from a fitted line against the target variable data points and tries to minimise it using gradient descent approach. The best fit line has the min least squares.

Post modelling analysis : Determine the coeffs for each variable to show how much each variable influences the target variable along with statistical significance

- Use the fitted model on test set to evaluate and predict

General info on Linear Regression

- Simple eq of y = mx +c where m is the slope and c is the intercept or $y = \beta_0 + \beta_1 * X_1 + \beta_2 * X_{2..}$
- Y —> Target variable (dependent variable vector)
- X_i —> Predictors or Features (matrix of independent variables)
- β_i —> β_0 is the intercept and $\beta_{i,2,3...}$ are the coefficients of predictors that determine the level of influence)
- We try to get the best values for intercept and coefficients by minimising the cost function which is the root mean square error between the predicted y value and true value of y
- Gradient descent is one way of reducing the cost function. We start with arbitrary values and move to a -ve gradient iteratively to get to the minima
- e_i —> error term = y_i - y pred and Residual Sum of Squares = sum of $(e_1^2+e_2^2+...)$ which is the cost function which at minimum value gives the best fit line.
- The final predicted equation tells us how much the value of y changes for per unit change in variables.
- Eg: ‘y’ changes by β_1 for 1 unit change in X_1 assuming other features remaining constant
- The strength of the regression line is given by R2 which is 1 - (RSS/TSS) where TSS, Total Sum of Squares is the difference between the true y values and the mean value $(y_i-y_{mean})^2$ for all i

2. Anscombe’s Quartet

- 4 data points that have almost the same summary statistics like mean, std deviation etc but have a different distribution form an Anscombe’s quartet.
- The difference in distribution is apparent only in the visual form and hence this lays additional emphasis on considering the visual representation of the data in addition to summary statistics

3. Pearson’s R

- Pearson’s R is the correlation coefficient which explains the relationship between given 2 variables.
- The values range from -1 to 1.
- While values close to 0 imply very low correlation (weak relationship) between the 2 variables, values close to 1 show a very strong relationship.

  +1 —> strong positive or direct relationship, values change in the same direction

  -1 —> strong negative or indirect relationship, values change in the opposite direction

  Eg:

  - Distance and time taken for a Bike trip will have + ve correlation. As distance increases, trip time increases

  - Avg Speed and time taken for a Bike trip will have -ve correlation. As the speed increases, the trip time decreases

- Pearson’s correlation coefficient between 2 variables X and Y is defined as

  (Covariance between X and Y) / (Std dev X)(Std dev Y)*

  It can be calculated using the numpy module and corrcoef method as shown below

  np.corrcoef(variable1,variable2)

4. Scaling

- Scaling is changing the values of variables in a multi linear regression model to ensure the right interpretability.
- Scaling is done to ensure the variables are in the same scale else there is a chance that higher weightage is given to variables that have higher values Eg: Bike trip price as a factor of distance in km, trip time in minutes and also count of active demand or customers
- Scaling is done using normalisation or standardisation
- In normalisation we use Min Max method where the min and max of the data points are used to convert all the values to a range of 0 to 1
- In standardisation we use the Mean and std deviation to convert the values to a range with mean '**μ**' and std dev 'σ'.
- *The key difference is that standardisation will have a range which is not necessarily between 0 and 1. Normalisation will ensure that the entire data is between 0 and 1, even if there are outliers.*

5. High VIF

- VIF indicates correlation between variables given by $1/(1-R_j^2)$.
- If the VIF is high > 10 or infinity it means there are variables in the model which have a very high correlation (R^2 close to 1)
- Sometimes constants in the regression models can have a high VIF, although it doesn't signify anything as the constants are not predictors
- In our model variables, had we included 'temp' and 'atemp' in the final model build, we could have seen high VIF for atemp. Hence it is recommended to drop the variables which have a high VIF, as these will not explain the target variable properly.
- As a general rule, it is suggested that we consider only the variables with VIF < 5. Variables with VIF > 10 to be dropped and between 5 and 10 to be investigated.

6. Q-Q plot

- Q-Q plot is a scatter plot used to compare the quantile values of 2 different data sets
- This plot helps understand if 2 data sets come from same population, if they have same distribution and scale
- If the 2 distributions are identical then the Q-Q plot would be similar to the straight line y=x
- We can understand the skewness, spread or variance of the distribution by checking the QQ plot
- Q-Q plot can be plotted even if 2 data sets have different sizes

Final Regression Eq

total_bookings = 0.2183+0.2307*(year)
+0.4960*(temperature)-0.0860*(holiday)-0.1406*(humidity)-0.1830*(windspeed)
+0.1180*(summer_flag)+0.0749*(fall_flag)
+0.1620*(winter_flag)-0.0522*(mist_flag)-0.2396*(light snow or light_rain)