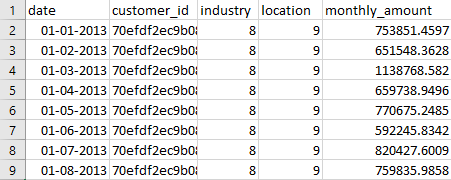
**Forecasting the transaction amount for December 2016**

**Business Understanding**  
We need to predict the transaction for December 2016. By predicting the transaction for December 2016, we can know which industry or location to prioritize for that month or it can help us understand how many employees to allot at a specific industry or at a certain location for that month to boost revenue. Meaning: Industries or Locations with higher predicted transactions for December 2016 can be given a higher priority by employing more people at those places. Whereas, industries and location with lower predicted transaction for December 2016 can be given lower priority by employing less people there. Doing so can make the company efficient and save the company money by not overemploying people at industries or locations where they won’t be needed. In the meantime, employing them at the right locations and industries where the transactions are predicted to be high.

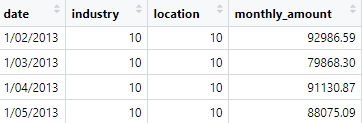
**Data Understanding & Data Preparation**We have a monthly data showing the transactions of 10 industries based at different locations. Below is the sample of the data:



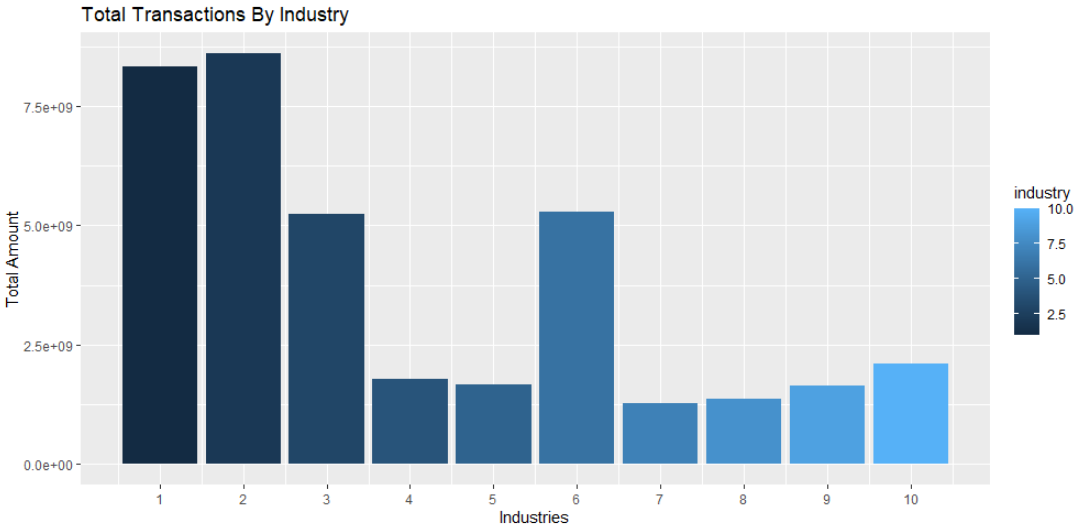
Here is a Data Dictionary to describe the different variables shown in the above dataset

|  |  |
| --- | --- |
| **Fields** | **Descriptions** |
| date | Stores the date of the first day of each month |
| customer\_id | Stores Unique customer ID’s |
| industry | Stores the code for 10 industries, ranging from 1 to 10 |
| location | Stores the code for 10 locations, ranging from 1 to 10 |
| monthly\_amount | Total transaction amount for a customer in given month |

The dataset as it is, is having no missing values. But certain dates seem to be missing for some industries. For example, For Industry 10 at Location 10, we have only 4 data entries as shown below.

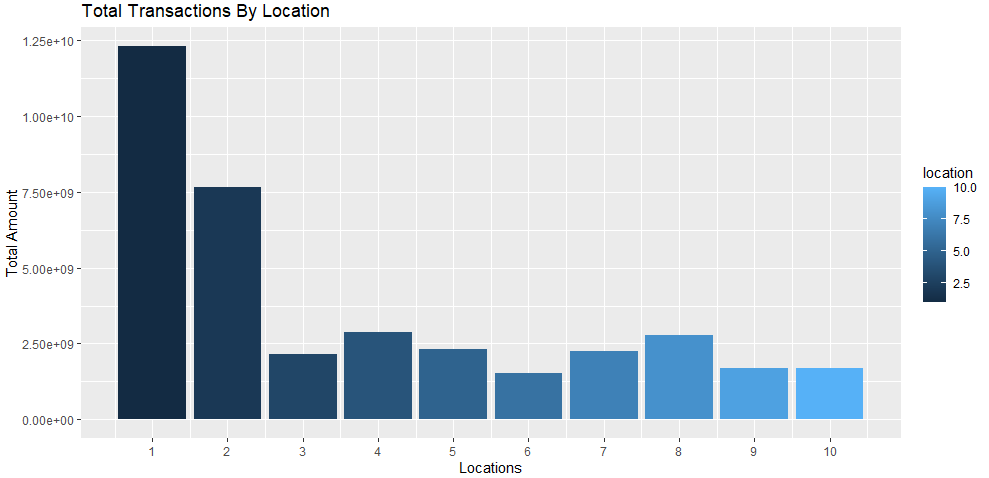


I then visualized the financial data to determine the total amount of transactions in every industry. Here Is a graph showing the Total transactions by each Industry:



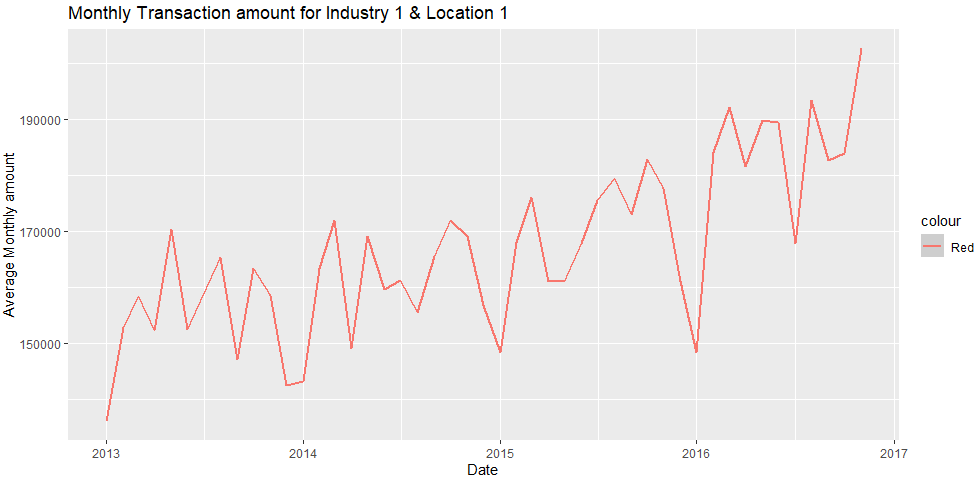
Evidently, Industries 2 & 1 have procured the highest total amount of transactions respectively, followed by Industries 6 & 3. Whereas, Industries 7 & 8 are industries that have the lowest total amount of transactions.

Thereafter, I then visualized the financial data to determine the total amount of transactions at each Locations. Here Is a graph showing the Total transactions by each Location:



Evidently, Location 1 & 2 have procured the highest total amount of transactions. Whereas Location 6, 9 and 10 are location have obtained the lowest total amount of transactions

In order to prepare the data: as data is not arranged by date, I first sort it by date and then create a new column in the dataset to ID each line. Furthermore, I created an aggregated data set using the fields date, industry and location, with a mean of monthly\_amount.

**Modelling**  
Firstly, we filtered industry 1 & Location 1 data from the aggregated data set, thereafter we created a Line plot illustrating the monthly\_amount for Industry 1 & Location 1. The Line plot is showcased below:  


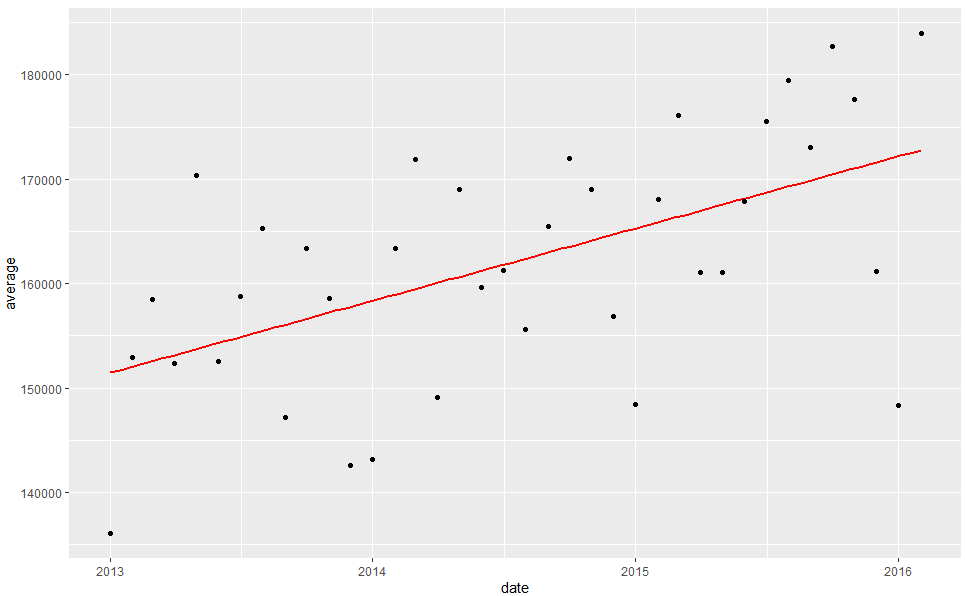
As seen in the graph the Monthly transaction amount is very high for the 1st quarter of the year, on the contrary the monthly transaction amount declines in the 4th quarter.

In order to predict the monthly transaction amount for December 2016 we will use a linear regression model. Linear regression model is a statistical analysis that shows relationship between two variables. Linear regression performs exceptionally well for linearly separable data and it uses the least square method.

As discussed earlier, monthly transaction amount is very high for the 1st quarter and it declines in the 4th quarter for Industry 1 at location 1. In order to account for this seasonal change in our data we drafted two variables namely: month and year in order to capture the seasonality of the data in our model.

Splitting the dataset: As there are 47 rows of aggregated monthly transaction data for Industry 1 at location 1. We allot 38 rows for training the model and 9 for testing the model.

Note: I had tried multiple other ways of splitting the data differently, But the difference between predicted value and the actual value was significantly high. Hence after trial and error, I selected 38 rows for training and 9 for testing as the difference between predicted and actual was minimum for this split

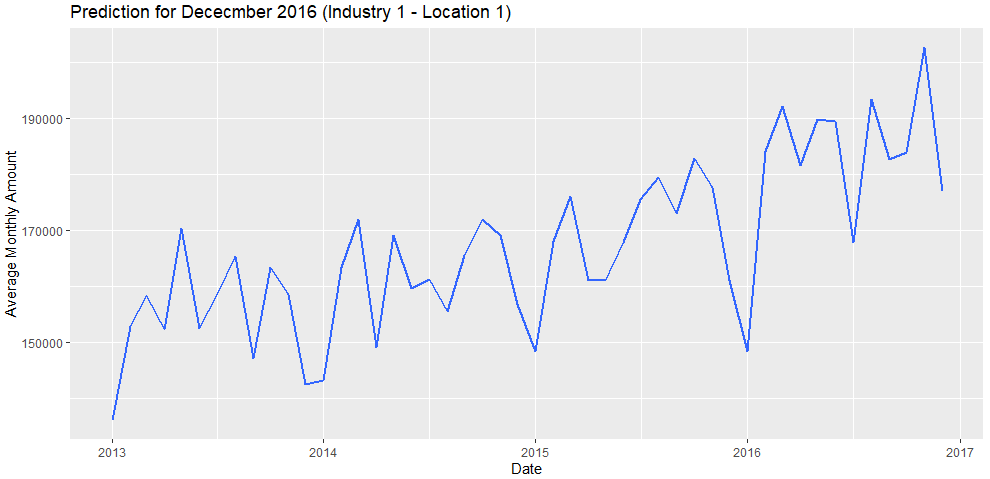
The predictor variables or the independent variables we used for our regression model was date, ID and month(in order to capture the seasonality) and the dependent variable was monthly amount. Median Residual from that model was 2334, Which is negligible. Hence, Model is a good fit. Furthermore, below is a linear regression model plot showing an upward trend:  


Note: The Regression line runs straight the middle of the data points  
Residuals is the measure of how far from the regression line, the data points are.

Upon applying the model to predict, it gave us the average predicted value of **176893.6** for December 2016 for Industry 1 at Location 1 as shown below:



As shown below in our prediction plot, December 2016 sales and the plot below also fits well/matches with the [monthly transaction amount plot for industry 1 at Location 1](#plot). The predicted value for December 2016 is less than that of the previous few months, hence seasonality was captured.



[Here](#actual) are the predicted values, actual values and difference between the predicted and actual attached in the appendix [below](#actual).  
Note: The difference percentage between actual values and predicted values is not much, It ranges from -14% to 6%. Model is a good fit for industry 1 at location 1.

**Evaluation and deployment**

I applied the linear regression model of industry 1 at location 1 for all the rest industries and locations in our data set. Even here by default we used only 9 rows for testing our data at the same time making sure that we have sufficient number of training rows. Upon applying the model, we created for industry 1 and location 1 on other industries we got very poor results. It is not feasible as it can only predict accurately less than 5% of data. [Here](#RMSE) is the result from our deployment showing the December 2016 predictions for all Industries and locations. [The result in the appendix](#RMSE) below also shows RMSE(Root mean square error). RMSE is the standard deviation of the residuals. The less the RMSE the better the model fits.

Hence, it seems that the model performed the worst for Industries 6 & 9. It performed the best for industries 1 and 10. On the other hand it performed the worst for locations 1 & 8. It performed best for locations 3 &5. Maybe the model might improve if we apply it to just Industries and remove the locations or vice versa. Perhaps, best thing to do would be to individually create linear models for each industry and location manually although that would be time consuming. Perhaps that would ensure us the best fitting model.

**Appendix**

**Below are the prediction values, Actual values, difference between Prediction and actual value and difference percentage for the 9 rows of training dataset of Industry 1 at Location 1:**

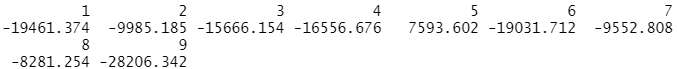
Predicted values

****

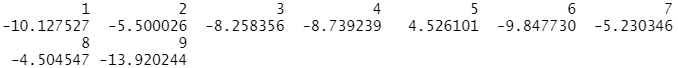
Actual values



Difference between Predicted value and Actual value



Difference percentage



**Dec 2016 Pre****diction and RMSE on all models**

Industry Location RMSE Dec 2016 Prediction

1 10 3 603.6323 79534.52

2 9 5 2012.4789 66600.21

3 10 1 4446.7008 49845.73

4 1 8 5538.6968 126570.66

5 10 4 7106.1359 106503.13

6 1 6 7387.6404 119469.94

7 1 2 8548.6526 210623.53

8 1 1 9515.8878 178543.72

9 1 3 9603.3477 205911.95

10 10 9 9608.8493 99452.73

11 1 5 9863.2697 173610.63

12 10 5 11517.5418 143048.51

13 1 9 11805.5751 198723.95

14 1 10 15489.9964 242128.47

15 1 4 15988.3260 208957.94

16 4 2 16095.2431 225127.74

17 7 7 17066.0262 226031.89

18 1 7 17740.0193 161432.82

19 5 5 17751.3792 241300.70

20 4 8 21034.0076 237613.89

21 4 9 21262.1287 225816.21

22 7 1 24685.4813 252164.27

23 2 2 24841.8438 482435.01

24 2 7 25802.9804 273000.87

25 2 10 26797.2850 284117.69

26 4 7 27077.5698 540174.60

27 4 1 28557.1759 491501.14

28 4 10 28749.5933 311294.54

29 2 1 29719.7977 429667.31

30 7 2 33023.1388 293742.06

31 5 2 34858.0487 429031.34

32 7 6 35929.7034 219105.71

33 7 8 36228.8703 303746.65

34 2 6 36700.1712 366288.29

35 4 3 37680.6488 322742.01

36 3 8 37833.7998 178052.35

37 5 7 39769.1310 329872.86

38 4 5 42716.2924 445966.52

39 4 4 43166.4768 419228.93

40 7 5 44955.1432 325303.54

41 8 4 47181.7160 354409.03

42 5 4 47942.3169 254968.32

43 2 4 48671.9104 255885.48

44 5 9 49200.0968 440365.08

45 2 8 49243.8068 500516.34

46 5 8 50651.0459 465697.13

47 7 4 53833.1084 428257.03

48 7 10 54213.4159 497701.79

49 5 3 57612.6727 299553.86

50 2 3 57701.5777 607024.75

51 10 2 59745.2694 238273.67

52 2 5 60035.4295 274250.46

53 3 2 60244.4948 688737.10

54 5 1 63162.4640 610234.60

55 3 7 67254.4225 648664.85

56 8 9 70051.8602 553241.26

57 2 9 70325.1443 460783.34

58 5 6 87166.2626 699754.56

59 7 3 87524.9314 313296.41

60 5 10 91846.4981 462057.03

61 7 9 92696.3108 779314.88

62 3 1 95393.2153 601300.03

63 8 2 96793.6409 760218.25

64 9 1 99543.2810 631064.75

65 3 6 103163.3940 125025.75

66 3 4 109780.5334 1027582.90

67 3 3 114278.7645 555142.52

68 10 7 115089.4627 299069.69

69 8 10 115965.6449 656832.62

70 8 6 139653.7829 1010321.22

71 3 5 146819.0369 332931.68

72 9 2 148446.3713 1256372.52

73 9 7 155797.6856 1521353.94

74 4 6 162835.4399 1139849.79

75 8 3 169513.3461 663966.23

76 9 3 179000.5605 1004181.55

77 8 8 179116.9772 437249.51

78 8 7 179337.4219 257994.28

79 3 9 181730.2677 375564.46

80 9 9 196809.2936 1137264.71

81 3 10 198078.4934 -199643.72

82 8 5 204982.4257 123212.52

83 8 1 284506.8298 1460039.43

84 9 6 297460.1852 1763920.96

85 9 10 322657.7149 1308952.43

86 9 4 539522.8534 -253357.89

87 10 8 2574934.2447 25676380.07

88 6 1 4372304.0117 35859656.20

89 6 2 NA NA

90 6 3 NA NA

91 6 4 NA NA

92 6 5 NA NA

93 6 6 NA NA

94 6 7 NA NA

95 6 8 NA NA

96 6 9 NA NA

97 6 10 NA NA

98 9 8 NA NA

99 10 6 NA NA

100 10 10 NA NA