First week submission

Introduction/Business Problem

Introduction

I selected a real life business problem for the submission of the Capstone project. I would like to state right at the beginning that this project does not rely or use the Fourquare APIs and data.

Still during the project I was using many of the tools and techniques that we have learned during the last couple of weeks of the Applied Data Science curriculum.

I worked with our company’s real data - that I have depersonalized carefully - to propose a possible solution for a monthly mobile telecommunication allowance to it’s employees based on historical reimbursements.

The basic idea is to assign people into groups – which will define the amount of the allowance – and based on that managers do not have to approve every single item on the mobile phone bills. Separate approval is only required if someone is spending above the allowance for business purposes (personal usage can not be reimbursed).

This approach can save time and money since people will try to fit within the approved packages and managers does not have to spend a lot of time to review their employees mobile expense claims.

Business Problem

During one of the usual management meetings we were discussing the cost / expense structure of our operations.

The team wanted to change certain aspects of the company’s compensation packages for employees / managers in line with the forecasted top line revenue and profit plan for 2019.

As part of the discussion we were debating on a fair amount that could be reimbursed by employees related to their usage of mobile telecommunication (voice and data) services on a monthly basis without prior approval.

After we covered a lot of ideas and opinions we have decided to use the data from the expense claim system and analyze the current situation as the first step.

We all agreed that this approach will help to better understand:

- How much the company pays per month for mobile telecommunication?
- Is there any trend related to the overall expenditure driven by the season?
- Which departments are the lowest and highest contributors?
- What is the average spend by department, is there any outlier?
- What is the typical amount spend by managers and practitioners (practitioners are the ones whose cost is charged to our Clients)?
- How we can group our employees based on their mobile expenditure?
- What would be a fair monthly allowance to these groups based on their current spending behavior?

Since I quickly realized that this is a segmentation problem and I have just learned about this recently on Coursera I volunteered to work together with our operations team and analyze the data.

A week after our initial discussion I have presented the findings on our next management meeting ….

Data description

Description of data

Data was extracted from the expense claim system by our operations team and given to me in .xls file.

The data contained the following information:

- Name of the person who submitted the claim;
- Date of submission;
- Department code of the person;
- Amount of the claim (in OMR);

To prepare the data for further analysis I decided to include additional information and change some of the data:

- I have changed the date of submission so it only contains the month;
- Added a flag to indicate whether the person is a manager or not;
- Added a flag to indicate whether the person is a practitioner or not;
- Changed the department code to the name of the department for easier analysis;
- Removed the person name from the analysis;

Once I completed the above mentioned modifications and added new data from various other data sources I have saved the file in a comma-separated values (.csv) file.

This .csv file is the one I used during my analysis.

Note

Usually the claim is submitted every month and it should contain the claimed value for the previous month. It happens occasionally that some people combine multiple months into a single reimbursement record and submit it all-together.

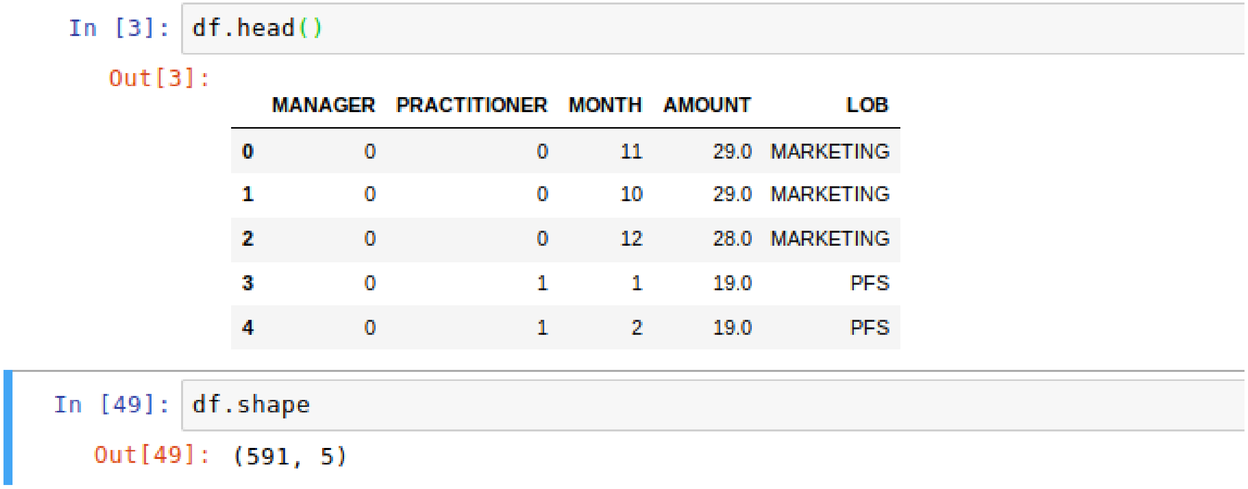
Although each month would show up as a separate line in the expense claim, since I did not have direct access to the expense claim system I was not able to verify whether the .xls – that was given to me by our operations team – showed these cases correctly or not (e.g. each and every claimed amount was assigned to the relevant month or they were combined and the month was picked randomly by operation).

I did not consider the above to create a huge distortion in the data and further analysis supported my theory about the same.

Basic data analysis

Once I loaded the .csv file into a data frame in my Notebook I have carried out a couple of basic analytical steps to better understand the nature of the data and the approach that can be used to solve the business problem.

The file has 5 columns and contains 591 records:



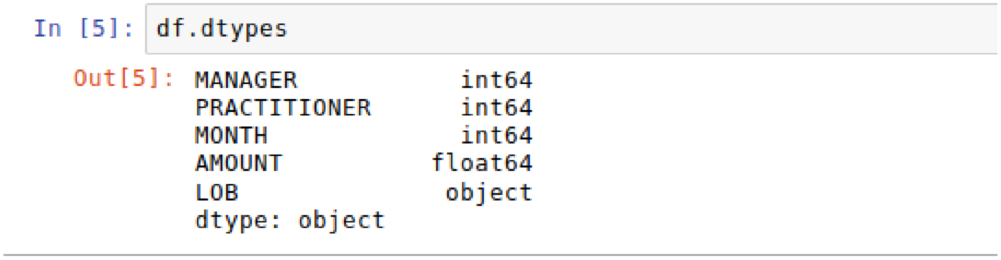
```
In [3]: df.head()
Out[3]:
```

	MANAGER	PRACTITIONER	MONTH	AMOUNT	LOB
0	0	0	11	29.0	MARKETING
1	0	0	10	29.0	MARKETING
2	0	0	12	28.0	MARKETING
3	0	1	1	19.0	PFS
4	0	1	2	19.0	PFS

```
In [49]: df.shape
Out[49]: (591, 5)
```

Illustration 1: Data frame head and shape

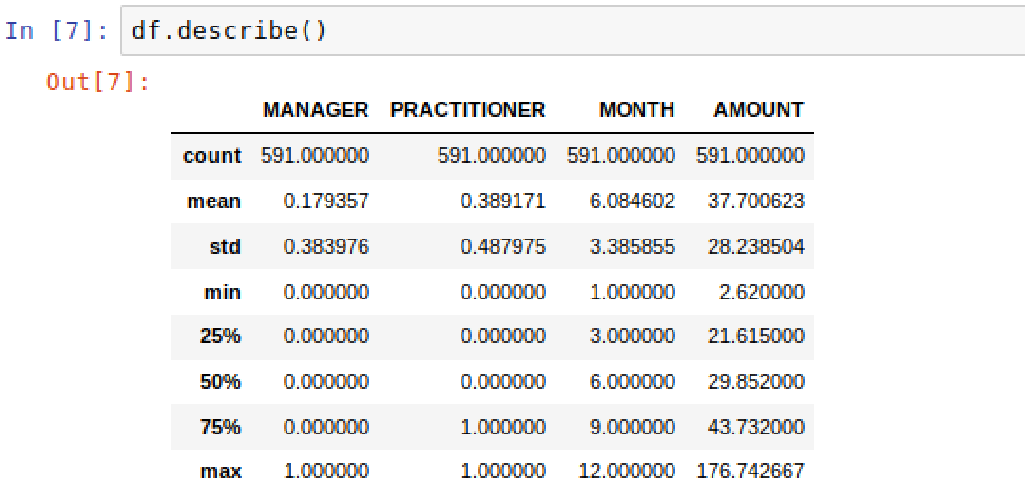
The columns has the following type:



```
In [5]: df.dtypes
Out[5]: MANAGER           int64
        PRACTITIONER      int64
        MONTH             int64
        AMOUNT          float64
        LOB              object
        dtype: object
```

Illustration 2: Data frame column data types

Basic analysis of the data frame shows that the average (mean) reimbursement amount is ~38 OMR:



```
In [7]: df.describe()
```

Out[7]:

	MANAGER	PRACTITIONER	MONTH	AMOUNT
count	591.000000	591.000000	591.000000	591.000000
mean	0.179357	0.389171	6.084602	37.700623
std	0.383976	0.487975	3.385855	28.238504
min	0.000000	0.000000	1.000000	2.620000
25%	0.000000	0.000000	3.000000	21.615000
50%	0.000000	0.000000	6.000000	29.852000
75%	0.000000	1.000000	9.000000	43.732000
max	1.000000	1.000000	12.000000	176.742667

Illustration 3: Basic analysis of the data frame

The basic analysis also showed that the data is equally distributed across the time period (it contained data from 2018 and early 2019).

Let us do a box plot to see whether our data supports this theory:

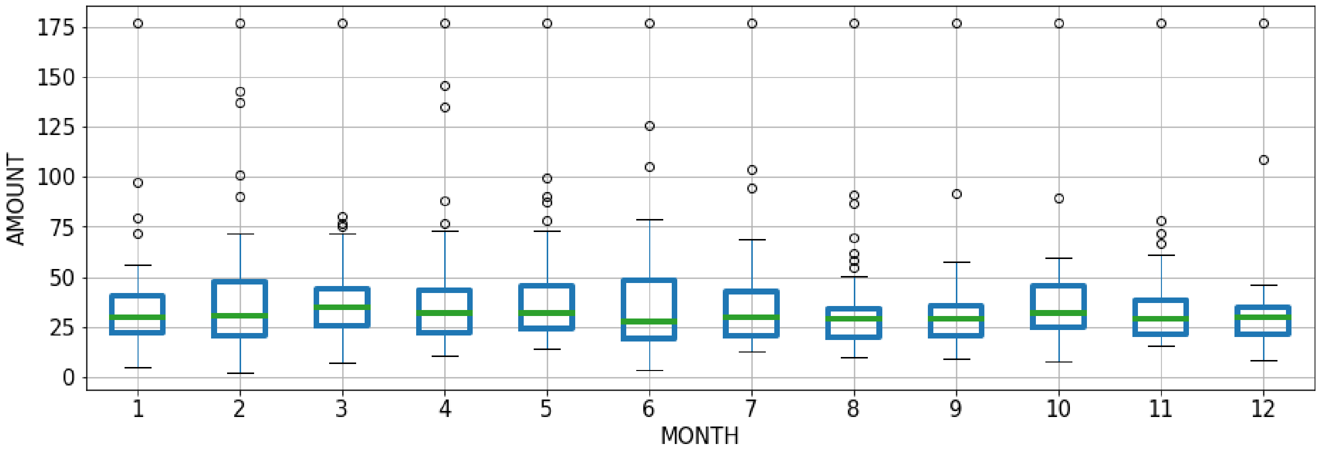
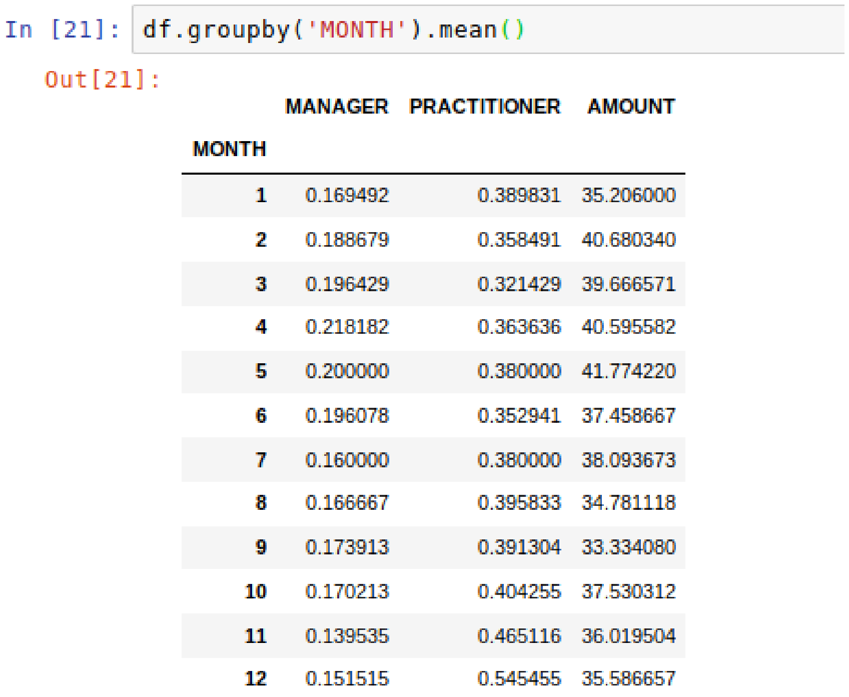


Illustration 4: Box plot showing the claimed amounts grouped by month

As you can see although there are outliers in each month the median for the quartiles are very close to each other. This suggests that the actual season or quarter does not have any impact on the claimed amount.

Let us do a group by on the data set to see it in numbers:



```
In [21]: df.groupby('MONTH').mean()
```

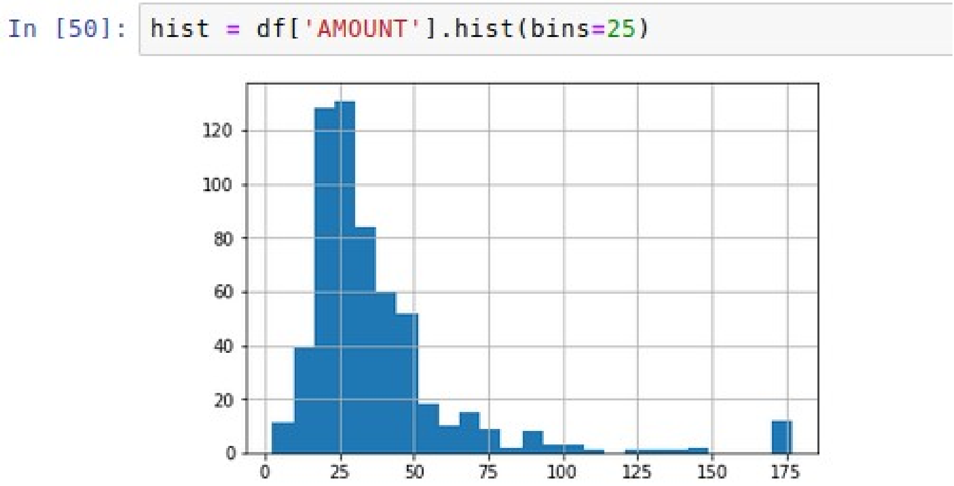
Out[21]:

	MANAGER	PRACTITIONER	AMOUNT
MONTH			
1	0.169492	0.389831	35.206000
2	0.188679	0.358491	40.680340
3	0.196429	0.321429	39.666571
4	0.218182	0.363636	40.595582
5	0.200000	0.380000	41.774220
6	0.196078	0.352941	37.458667
7	0.160000	0.380000	38.093673
8	0.166667	0.395833	34.781118
9	0.173913	0.391304	33.334080
10	0.170213	0.404255	37.530312
11	0.139535	0.465116	36.019504
12	0.151515	0.545455	35.586657

Illustration 5: Data frame grouped by MONTH and aggregated

None of the months seems to deviate too much from the average monthly claim (as we have seen above it is around ~38 OMR).

All-right, as a last basic data analysis let us see how the amounts are distributed. The following histogram shows that most of the amounts are between 15 and 35 OMR.



```
In [50]: hist = df['AMOUNT'].hist(bins=25)
```

Illustration 6: Histogram of amounts claimed

Second week submission

Methodology

The methodology section describes the exploratory data analysis and the inferential statistical testing performed, and the machine learning approach used for the analysis of the data set.

As a first step I have checked the distribution of the claimed amounts based on the MANAGER flag. The below figure shows the results (0 – Non manager, 1 - Manager).

The data set contained 106 expense claims submission for managers and 485 for non managers.

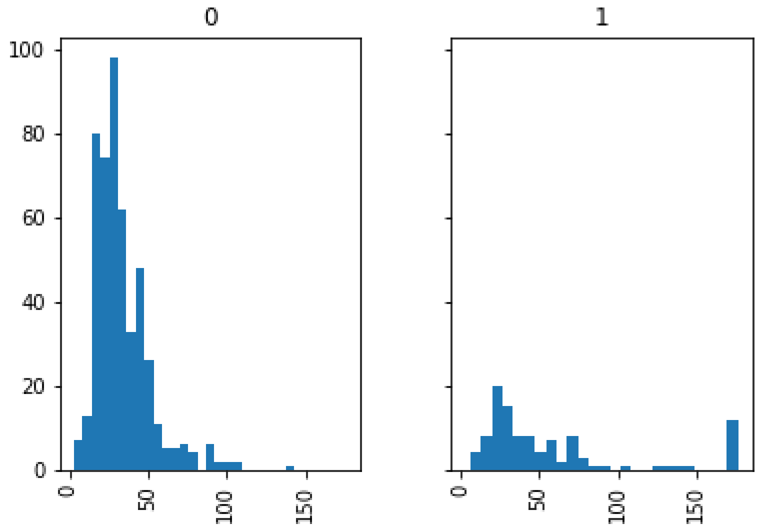


Illustration 7: Histogram of amounts claimed based on the manager categorical variable in the data set

Then I checked the same for the PRACTITIONER categorical variable as well. The below figure shows the results (0 – Non practitioner, 1 – Practitioner).

The data set contained 230 expense claims submission for practitioners and 361 for non practitioners.

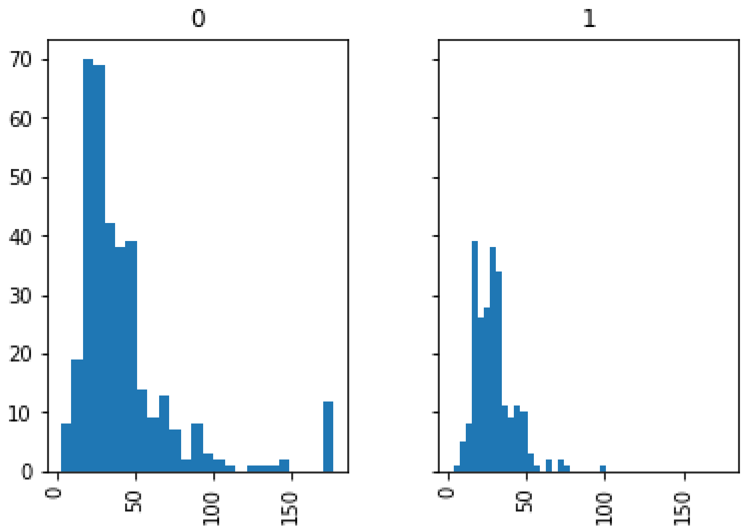
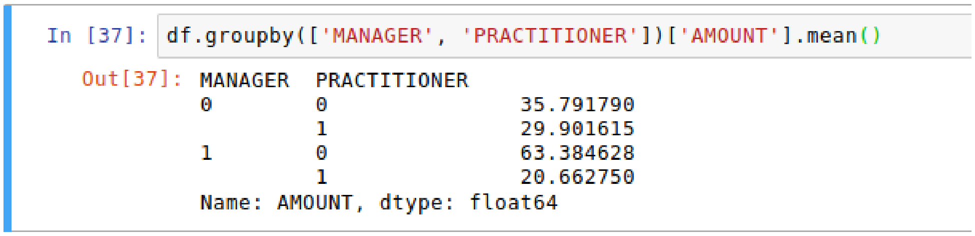


Illustration 8: Histogram of amounts claimed based on the practitioner categorical variable in the data set

For the shake of completeness I checked the mean of the submission by these two features (MANAGER, PRACTITIONER) for the whole data set. I will use it later to see how much it differs from the actual results of clustering.



```
In [37]: df.groupby(['MANAGER', 'PRACTITIONER'])['AMOUNT'].mean()

Out[37]: MANAGER  PRACTITIONER
         0        0               35.791790
                  1               29.901615
         1        0               63.384628
                  1               20.662750
         Name: AMOUNT, dtype: float64
```

Illustration 9: Mean value of claimed amounts for managers and practitioners

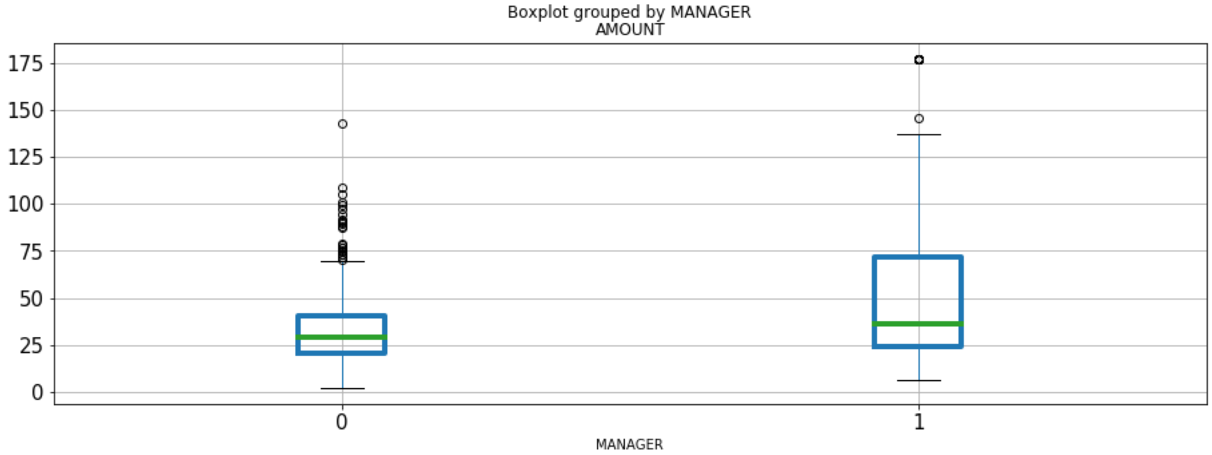


Illustration 10: Box plot showing the claimed amounts for managers and non-managers

I have also created a density plot that you can check in the notebook attached to the final submission.

The box plot below shows the distribution of the claimed amounts across the time period for each month.

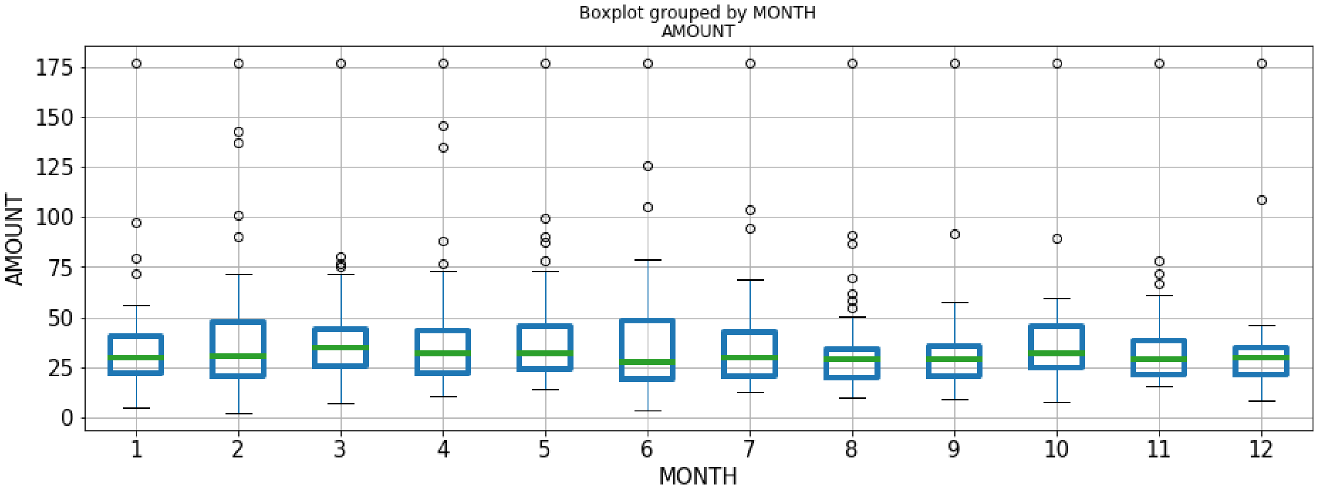


Illustration 11: Box plot showing the claimed amounts grouped by month

I have also run an analysis to check the distribution of the claimed amounts across the various line of businesses (LOB).

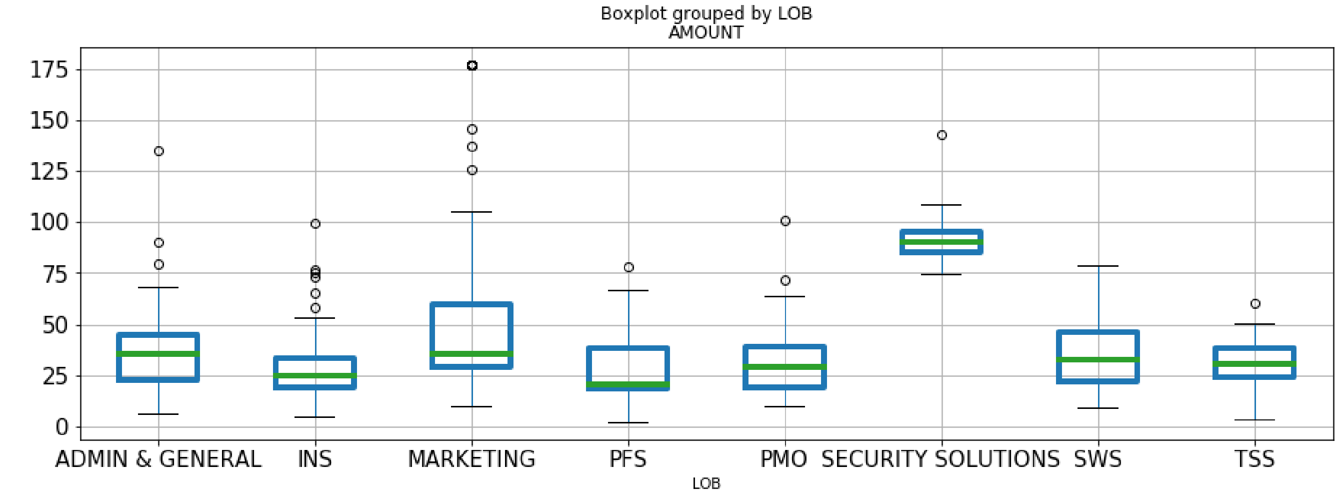


Illustration 12: Box plot showing the distribution of amounts claimed by LOB

As the above figures show most of the months and LOBs are in the same range. The Security Solutions line of business seems to be an exception which might need further investigation.

Linear regression

As a first step I have checked whether there is any linear relationship exists between the amounts claimed and the other features in the data set. I have used SLR and MLR techniques as well with a randomly selected training and test data set.

The intuition based on the preliminary statistical checking was that no such relationship should exist.

First I have checked whether the LOBs can be related to the monthly amount or not. I set the predictors to the columns representing the LOB and the target variable to AMOUNT.

This MLR model give a variance score of 0.14. This is just a little bit better than the fit of the mean line therefore I moved on to check whether the LOB separately and the AMOUNT are linearly related or not.

Based on the results (see the details in the Notebook attached to the final submission) none of the LOB is linearly related to the AMOUNT (in all cases the variance score was close to zero or even a negative number meaning that the fitted line is worse than the mean).

K-Means

I have selected the K-Means clustering algorithm to create a segmentation for the expense claims based on the features in the data set. This is an un-supervised machine learning algorithm that can be used for such a purpose.

To identify the best k value (number of clusters) I have generated a list of the sum of the squared distances with different k values. This is a common approach that can be used to identify the best possible k value.

The following figure shows the results of the "Elbow method". As you can see on the figure the optimal value is at k = 8 but also k = 2 gives a significant drop in the sum of the squared distances.

In my evaluation I will use both values.

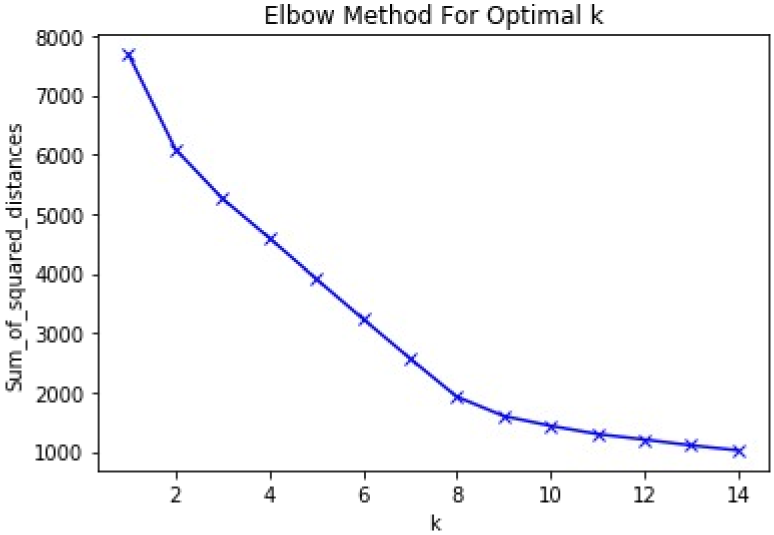


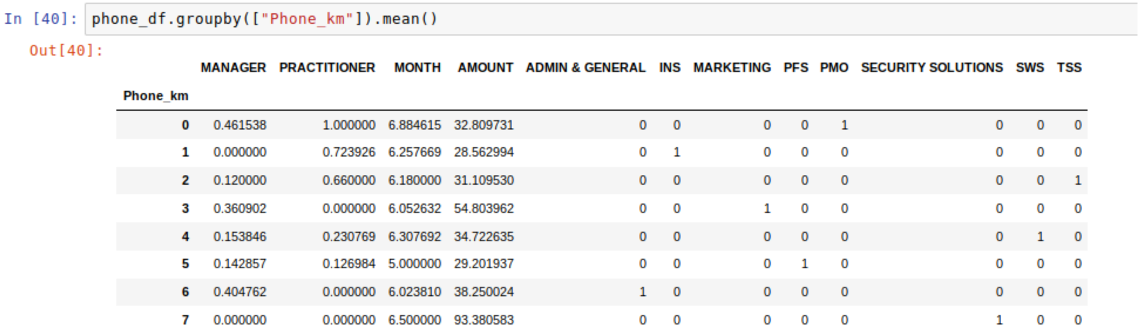
Illustration 13: Elbow point showing the optimal number (k) of clusters for analysis

Results

I run the K-Means algorithm with two possible k values based on the "Elbow method":

- I have checked the segments generated with k = 8 and
- also with k = 2

In case of k = 8 the K-Means machine learning algorithm provided the following clusters:

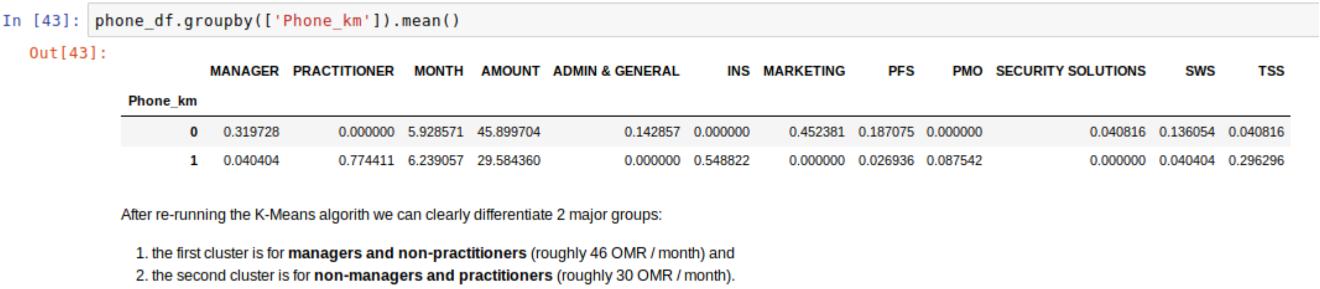


```
In [40]: phone_df.groupby(["Phone_km"]).mean()
Out[40]:
```

Phone_km	MANAGER	PRACTITIONER	MONTH	AMOUNT	ADMIN & GENERAL	INS	MARKETING	PFS	PMO	SECURITY SOLUTIONS	SWS	TSS
0	0.461538	1.000000	6.884615	32.809731	0	0	0	0	1	0	0	0
1	0.000000	0.723926	6.257669	28.562994	0	1	0	0	0	0	0	0
2	0.120000	0.660000	6.180000	31.109530	0	0	0	0	0	0	0	1
3	0.360902	0.000000	6.052632	54.803962	0	0	1	0	0	0	0	0
4	0.153846	0.230769	6.307692	34.722635	0	0	0	0	0	0	1	0
5	0.142857	0.126984	5.000000	29.201937	0	0	0	1	0	0	0	0
6	0.404762	0.000000	6.023810	38.250024	1	0	0	0	0	0	0	0
7	0.000000	0.000000	6.500000	93.380583	0	0	0	0	0	1	0	0

Illustration 14: K-Means clusters created on the data set with k = 8

This clustering basically creates a segment for each and every line of business. I will talk about this in Conclusion chapter of this report.



```
In [43]: phone_df.groupby(['Phone_km']).mean()
Out[43]:
```

Phone_km	MANAGER	PRACTITIONER	MONTH	AMOUNT	ADMIN & GENERAL	INS	MARKETING	PFS	PMO	SECURITY SOLUTIONS	SWS	TSS
0	0.319728	0.000000	5.928571	45.899704	0.142857	0.000000	0.452381	0.187075	0.000000	0.040816	0.136054	0.040816
1	0.040404	0.774411	6.239057	29.584360	0.000000	0.548822	0.000000	0.026936	0.087542	0.000000	0.040404	0.296296

After re-running the K-Means algorith we can clearly differentiate 2 major groups:

1. the first cluster is for **managers and non-practitioners** (roughly 46 OMR / month) and
2. the second cluster is for **non-managers and practitioners** (roughly 30 OMR / month).

Illustration 15: K-Means clusters created on the data set with k = 2

This clustering can be used to create 2 basic segments that can be easily implemented without creating a tension between the business units.

Please note how much this is different from the mean values of the data set when grouped by managers and practitioners (see Illustration 9 earlier in the Report).

The following histogram shows the 2 clusters color coded:

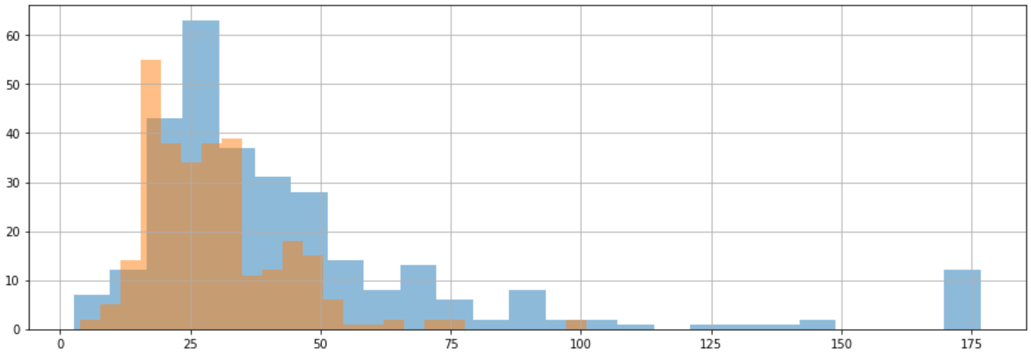


Illustration 16: Color coded clusters on a histogram

The total amount reimbursed for the first cluster (managers and non-practitioners) shown in blue color is 8786 OMR.

The total amount reimbursed for the second cluster (non-managers and practitioners) shown in orange color is 13.494 OMR.

Discussion

Running a regression on the data set showed that there is no linear relationship between the predictors and the target feature (AMOUNT).

It would be an interesting exercise to enrich the original data set with additional features that can help to build such a relationship.

With this in place a machine learning technique could be used to see whether a certain reimbursement is in line with the prediction or it is off.

In such a case further investigation could be triggered supporting the work of internal business control.

An other idea would be to split the reimbursement amounts and show how much is related to:

- National calls
- International calls
- SMS
- Data (with and without OTT)

Such additional data would help to better understand the usage pattern and to build a better segmentation model.

Conclusion

Running the clustering algorithm on the data set resulted in 2 different recommendations for the original business problem:

- Either create a separate segment for each line of business or
- create a segment for managers and non-practitioners and non-managers and practitioners.

Although the first approach might be a better one from a machine learning perspective (see the figure in the K-Means chapter of the Methodology section of this report) it is not easy to implement in a day-to-day operation. **It would also create potential tensions between the different business units.**

The second option provides an approach that can be easily rolled-out and maintained and could be accepted by everybody in the organization.

So let us analyze the possible impact of implementing the recommended clustering for employees.

The following table shows the number of claims per category:

	Practitioner	Non-practitioner
Manager	12	94
Non-manager	218	267

What would be the overall amount of expense claims having in place the recommended clustering?

Our first cluster is for managers and non-practitioners with a monthly allowance of 46 OMR (everything else on top of this amount would require an extra approval).

- This would result in 94x46 + 267x46 = 16.606 OMR

The second cluster is for non-managers and practitioners with a monthly allowance of 30 OMR (everything on top of this amount would require an extra approval).

- This would result in 12x30 + 218x30 = 6.900 OMR

Assuming that this segmentation and monthly allowance would have been in place this would result in 23.506 OMR overall reimbursement for the given time frame (I do not count with any extra expenditure on top).

This is 5.5% higher than the overall amount reimbursed in the original data set (22.281 OMR).

Considering the **1. reduced time required for administration** and also the **2. factor of extra approval required** to go beyond the approved monthly amount I see a great opportunity to better manage the telecommunication expenses for the company.

I would like to conclude my report with this final thought.