Faculty of Engineering
Cairo University

# Data Science

## Project Report Team 14

Supervised by: Dr.Dina Elreedy



| Name | Section | B.N |
|---|---|---|
| Ahmed Ayman | 1 | 1 |
| Omar Ahmed Mohamed | 2 | 4 |
| Ammar Mohamed Sobhi | 2 | 3 |
| Yousef Atef Tawfik | 2 | 40 |

## Contribution

| |
|---|
| Ahmed Ayman: Q4, Q6, Q8 |
| Omar Ahmed Mohamed: Q9, Q10 |
| Ammar Mohamed Sobhi: Q3, Q5, Q7 |
| Yousef Atef Tawfik: Q1, Q2, help with Q10 |

## Problem and Dataset Description

Our project covers the problem of Medical Representatives visits to Doctors. We will try to analyse their performance and answer the question of how to increase their productivity.
The dataset is retrieved from a real database of an application called Tacitapp that multiple Pharmaceutical Companies use in The Middle East and North Africa region.

The dataset contains data about multiple companies each company has dozens of medical reps that made thousands of visits in the last 5 years, and contains data about more than 60000 doctors.
Each visit contains data about the rep, the doctor, the product presented, the exact location where the visit was made, the total duration, the time of the visit, the duration of each slide in the presentation and any objections or remarks by the doctor.

## Q1) What is the average duration of a medical rep's visit to a doctor for a specific product?

## 1. Stating and refining the question.

**Epicycle**:
- Expectations: The question is answerable and useful.
- Collection: Looking at the dataset and collecting information about the company objectives.
  - conclusion: indeed time spent in the visit can be a good estimate of how good the rep can deliver the information, using the duration for every visit in the dataset we can get the average
- Match: indeed expectations and data collection match.

## 2. Exploratory data analysis.

**Epicycle**:
- Expectations: we should have a dataset of representatives visits with the duration of the visit and the product, we should not display any sensitive information of the users, we should not have missing values as the app forces the user to enter the complete information, however there could be some outliers due to users' mistakes.
- Collection: Exploring and vislualizing the dataset, looking for missing values and outliers, and also looking for any sensitive information that should not be displayed.
  - conclusion: indeed we have a dataset of representatives' visits with the duration of the visit and the product, we carefully drop any sensitive information before beginning the analysis, we have no missing values, however we have some outliers that we should take care of.
- Match: indeed expectations and data collection match.

**Missing values:**
we can see that there are no missing values in the dataset, and that is mainly because the system does not allow the rep to submit the visit if he did not fill all the required fields.

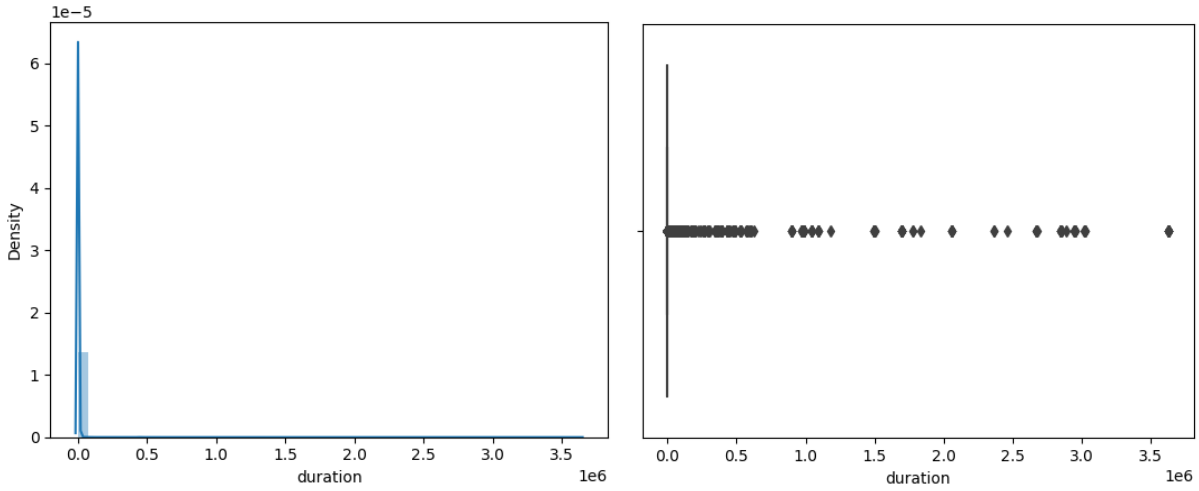
| | Total | Percent |
|---|---|---|
| representative_id | 0 | 0.0 |
| product_id | 0 | 0.0 |
| duration | 0 | 0.0 |

**Outliers:**

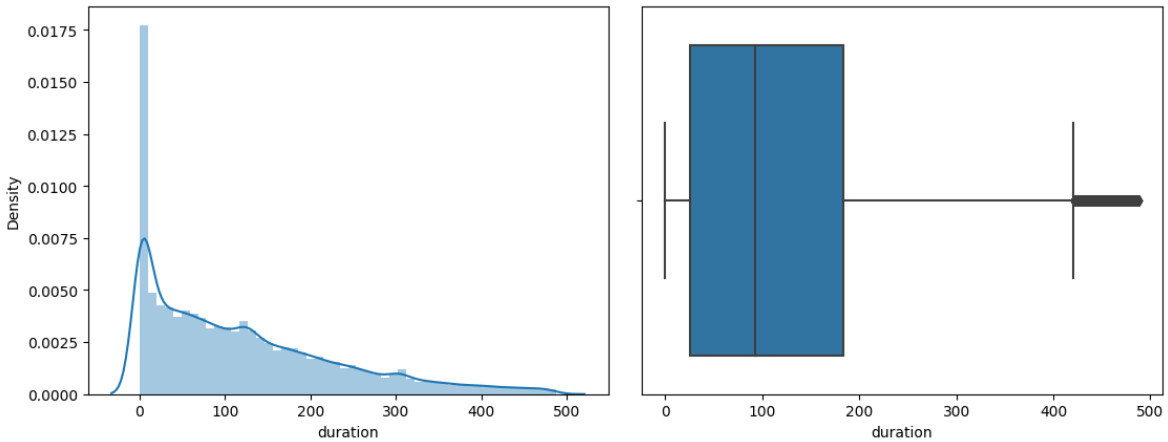
| | duration |
|---|---|
| count | 1.205210e+05 |
| mean | 2.288388e+03 |
| std | 6.607555e+04 |
| min | -1.603604e+06 |
| 25% | 3.100000e+01 |
| 50% | 1.050000e+02 |
| 75% | 2.140000e+02 |
| max | 3.631316e+06 |

We notice that the duration column has a minimum value less than 0, which is not possible, However the number of rows with negative duration is very small compared to the total number of rows (21 out of 120521) so we will drop these rows since they will not affect our analysis.

From histogram, skewness, kurtosis and boxplots we notice that: The distribution is highly skewed to the right; it indicates that there are a large number of extreme values on the right side of the distribution, with the majority of values concentrated on the left side, the distribution has heavy tails and a very peaked or sharp peak compared to a normal distribution and that suggests the presence of many outliers or extreme values in the dataset.




Indeed we seem to have some outliers, let's use the IQR method to detect and remove them, we get these histogram and boxplot:



## 3. Build a model.

however we don't need a model to answer our question, we can simply calculate the average of the duration column so we will do that:

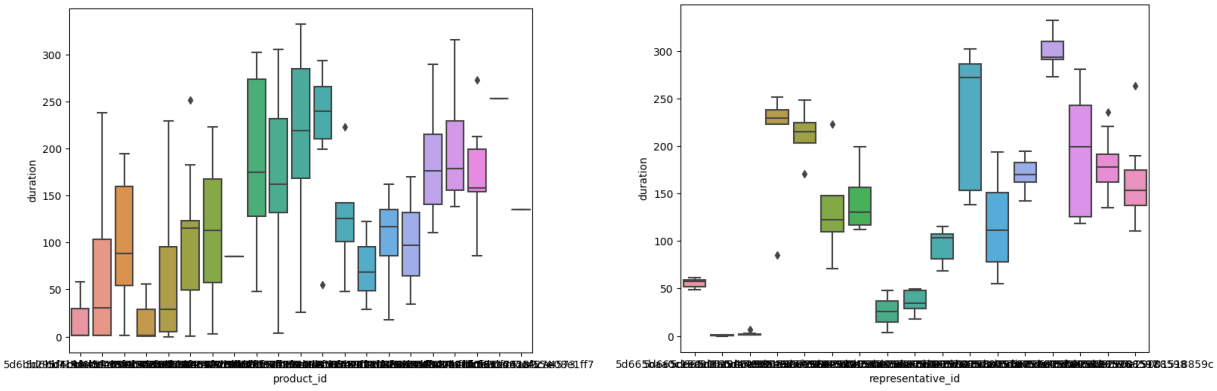
**Epicycle**:
- Expectations: After preprocessing and visualizing the data, we should be able to answer the question by grouping on the representative and product columns and calculating the average of the duration column.
- Collection: Indeed we managed to answer the question.
- Match: Indeed expectations and data collection match.

## 4. Interpret the results:

**Epicycle:**
- Expectations: Useful insights and interpretations, we should be able to indicate the distribution of the duration for each product, hence can be used to indicate the performance of the rep for each product.
- Collection: looking at the last query result we can see that the average duration of each product varies from a representative to another and some durations deviate from the average by a large amount, so we can conclude that which means that some representatives are better than others in delivering the information.
- Match: Indeed expectations and data collection match.

## 5. Communicate the results:

**Epicycle**:
- Expectations: The results should be communicated in a clear and understandable way.
- Collection: Indeed we used boxplots to show the distribution of the duration for each product.
- Match: Indeed expectations and data collection match.



## Q2) Are Doctors assigned to rep close to each other?

# 1. Stating and refining the question.

**Epicycle:**
- Expectations: The question is answerable and useful.
- Collection: Looking at the dataset and collecting information about the company objectives.
    - conclusion: Doctors assigned to rep are better to be close to increase the performance of the rep, using the geolocations of each doctor we can know how close are the assigned doctors, However a lot of companies still do not use the new geolocation feature hence we won't be able to calculate the exact distance between assigned doctors for some companies.
- Match: Mismatch between the data and the question, the question is not answerable with the given data.

We can edit the question to be "Are Doctors assigned to a representative in the same area?”

**Epicycle:**
- Expectations: The question is answerable and useful
- Collection: Looking at the dataset and collecting information about the company objectives.
    - conclusion: Doctors assigned to rep are better to be close to increase the performance of the rep, this time we can use the location of each doctor instead of the geolocation to know how many different areas the assigned doctors are in.
- Match: Match between the data and the question, the question is answerable and useful.

# 2. Exploratory data analysis.

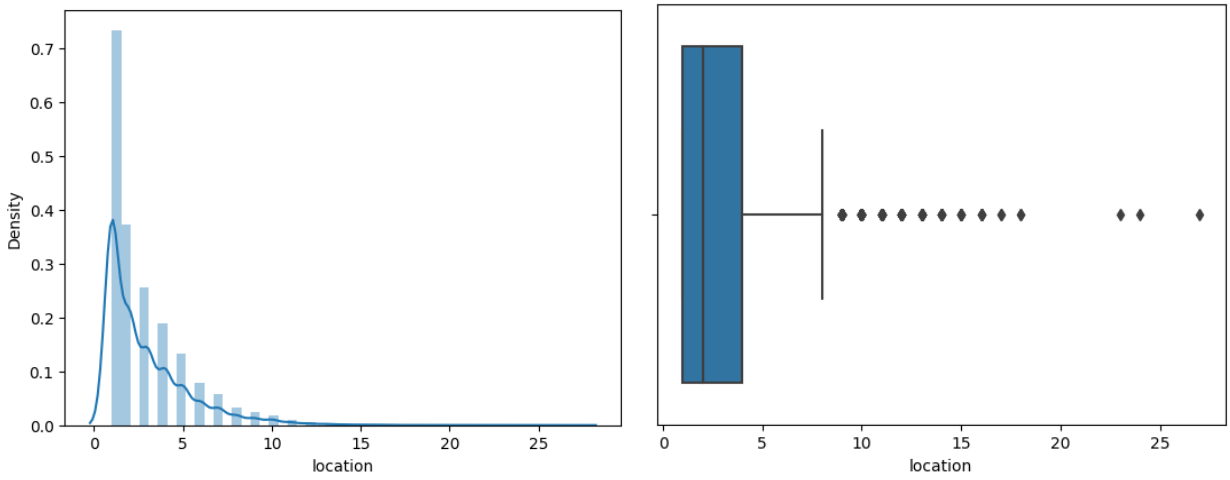
**Epicycle**:
- Expectations: we should have a dataset of representatives’ visits with the location of the visit, we should not display any sensitive information of the users, we should not have missing values as the app forces the user to enter the complete information, however there could be some outliers due to users' mistakes.
- Collection: Exploring and vislualizing the dataset, looking for missing values and outliers, and also looking for any sensitive information that should not be displayed.
    - conclusion: indeed we have a dataset of representatives’ visits with the location of the visit, we carefully drop any sensitive information before beginning the analysis, we have no missing values, however we have some outliers that we should take care of.
- Match: indeed expectations and data collection match.

**Missing values:**
we can see that there are no missing values in the dataset, and that is mainly because the system does not allow the rep to submit the visit if he did not fill all the required fields.

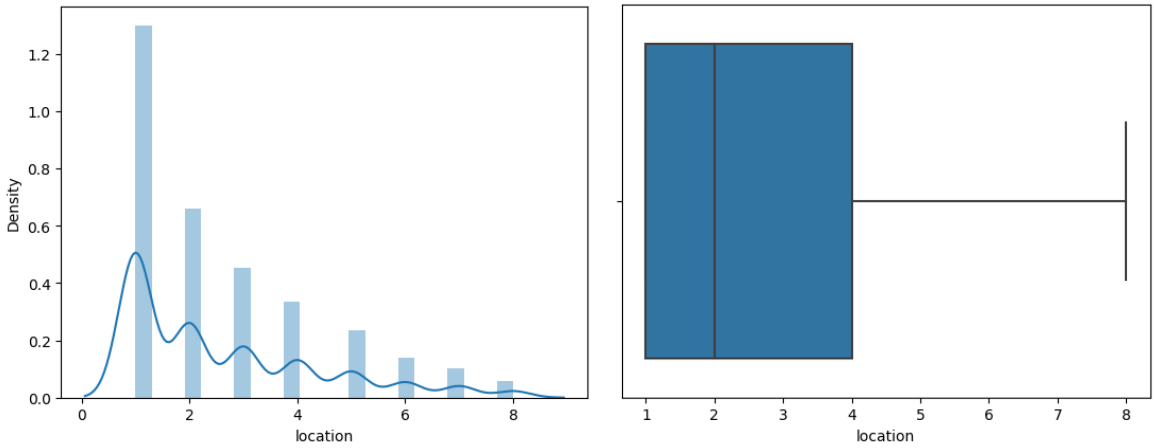
| | Total | Percent |
|---|---|---|
| representative_id | 0 | 0.0 |
| country | 0 | 0.0 |
| province | 0 | 0.0 |
| city | 0 | 0.0 |
| date | 0 | 0.0 |

**Outliers:**

From histogram, skewness, kurtosis and box plots we notice that:
There are definitely some outliers in the data, for example a representative cannot go to 27 different places in the same day.




Indeed we seem to have some outliers, let's use the IQR method to detect and remove them, we get these histogram and boxplot:




## 3. Build a model.

However, we don't need a model to answer our question, we can simply calculate the average number of different locations assigned to each representative per day.

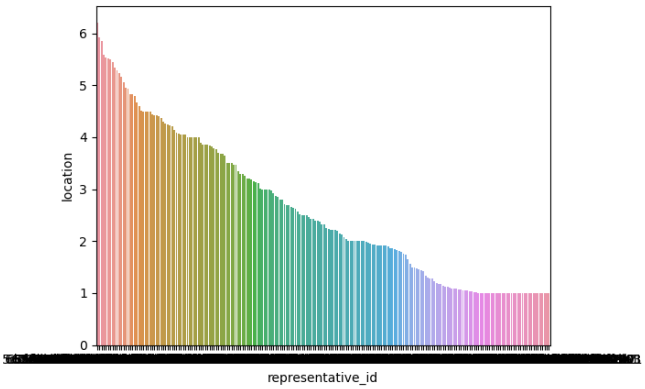
**Epicycle**:
- Expectations: After preprocessing and visualizing the data, we should be able to answer the question by grouping on the representative and day columns and calculate the average of the location column.
- Collection: Indeed we managed to answer the question.
- Match: Indeed expectations and data collection match.

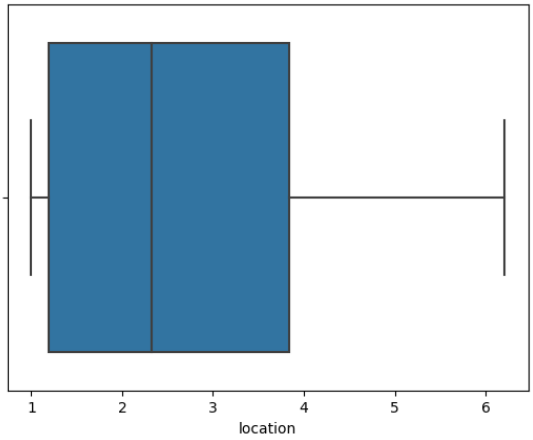
## 4. Interpret the results:

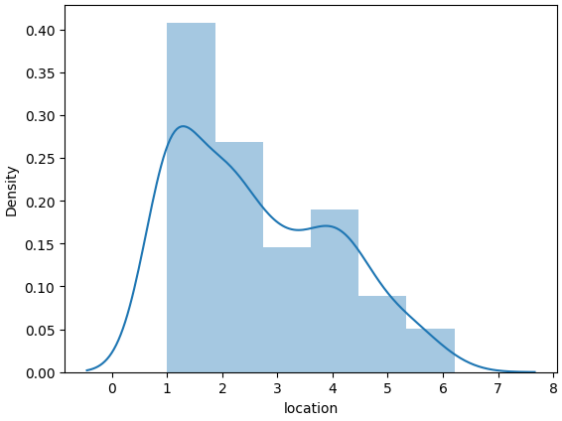
**Epicycle:**
- Expectations: Useful insights and interpretations of the results, also we should be able to get the distribution of the average number of different locations assigned to each representative per day.
- Collection: Indeed we can get the distribution and visualize it, and we can see that the average number of different locations visited by a representative in a single day varies from one representative to another.
So we can conclude that some representatives have to travel more than others and that can be exhausting, time consuming and can affect the performance of the representative.
- Match: Indeed expectations and data collection match.

## 5. Communicate the results:

**Epicycle**:
- Expectations: The results should be communicated in a clear and concise way, we should be able to communicate the results in a way that is understandable by the stakeholders.
- Collection: Indeed we can use boxplot and barplot to visualize the distribution and communicate the results.
- Match: Indeed expectations and data collection match.



## Q3) Is there a deviation between visits locations and actual locations of the doctors?

### 1- Stating and Refining the question

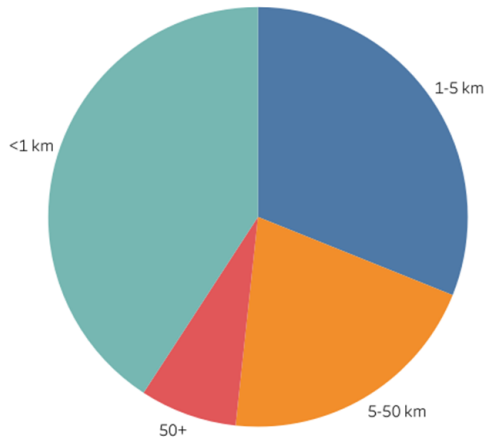
- **Expectation**: Get enough data to determine if there are fake visits or check accuracy of location service, We expect data of visits to contain geographical coordinates of each visit and the doctor himself to have his coordinates recorded.
- **Collecting info**: We check the dataset of visits and doctors
- **Comparing Data and expectations**: Our expectations match as we find the data we need and the question will be very useful for both software developers and the pharma companies to track fake visits.

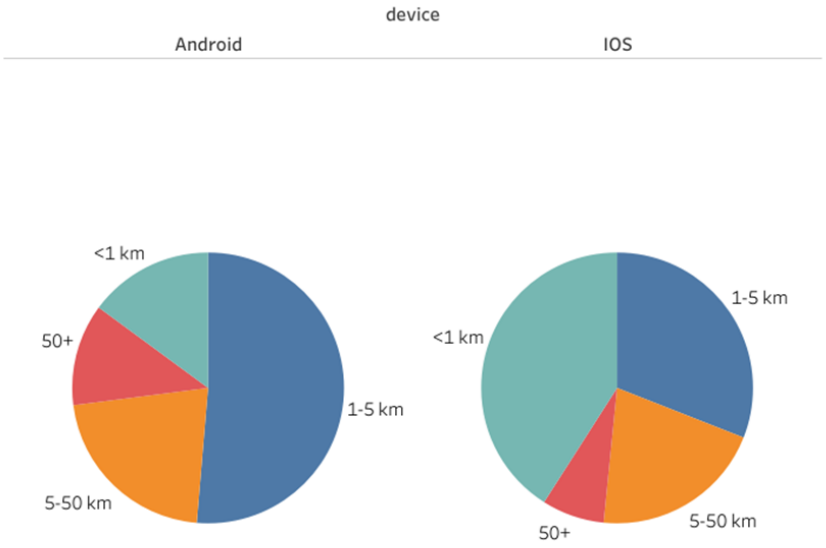
### 2- Explore Data

- **Expectation**: We expect to find enough data of visits where locations are correctly recorded and enough doctors where their locations are recorded.
- **Collecting info**: We find that only Doctors in Egypt have their location recorded (more than 3600 doctors) and we find more than 13400 visit records in Egypt with recorded location.
- **Comparing Data and expectations**: Since relevant data only exist in Egypt we refine our question to check deviation only in Egypt.

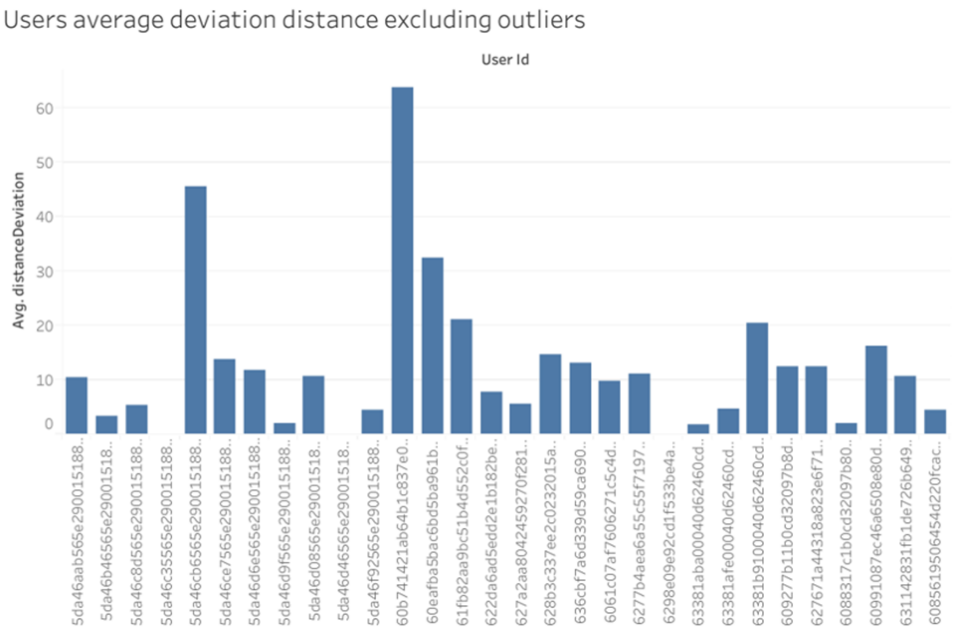
### 3- Model

- **Expectation**: Build visualisation on Tableau that answers our question
- **Collecting info**: We have the info we need.
- **Comparing Data and expectations**: Visualisation helped us answer our questions to check deviations in data, but it leads to further analysis






Some more analysis was needed to check deviations from user perspectives



## 4- Interpreting results

- **Expectation**: Clear interpretation that tells us the source of deviations if exists.
- **Collecting Info**: The result visualisations
- **Comparing Data and expectations**: Interpretations from the visualisations were useful, Data shows shows that only less than 50% of visits have acceptable deviation (less than 1km), Also Android devices have bigger problems than iOS devices in detecting locations, but it also shows most users have high deviations.

## 5- Communicating results

- **Expectation**: Clear communication that leads to clear steps on how to solve the problem
- **Comparing Data and expectations**: Our interpretation shows that further investigation needed whether Doctor's locations are actually correct or it might be outdated, it could suggest for app developers also to add a feature that allows multiple locations for the same doctor, and also location service needs to be rechecked by developers

## Q4) What is the distribution of doctors' locations?

# 1- Stating and Refining the question

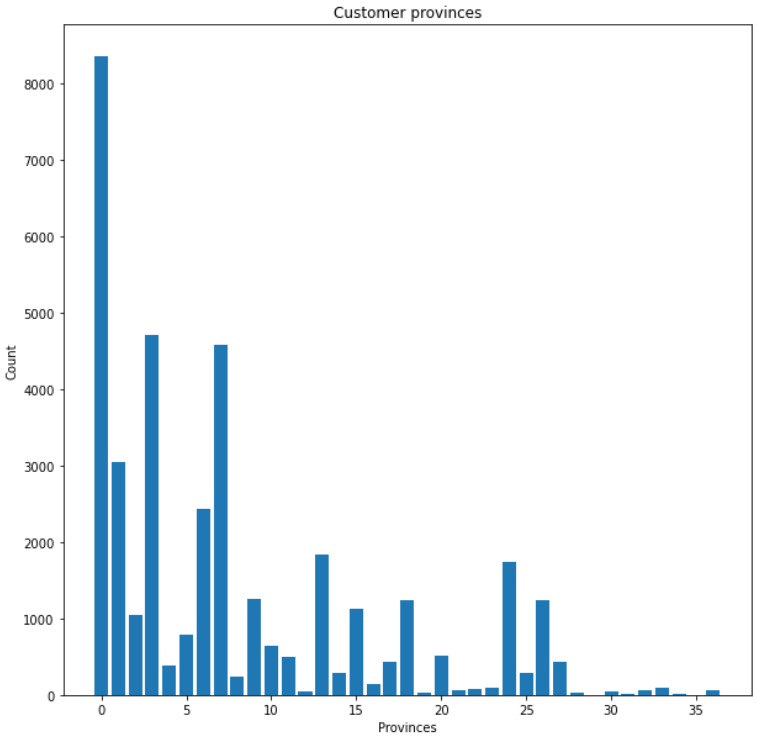
- **Expectation**: For doctor's locations distribution, we expect to find the location object for each customer that contains country, province and city.
- **Collecting info**: We check the importance of the question and ability to answer it.
- **Comparing Data and expectations**: Our expectations match as we find the data we need and the question will be very useful for pharma companies.

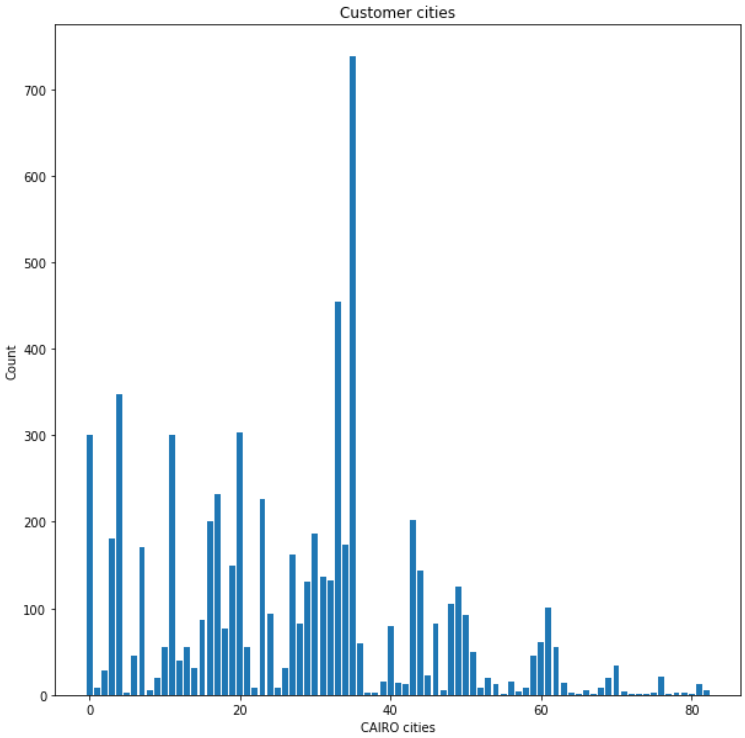
# 2- Explore Data

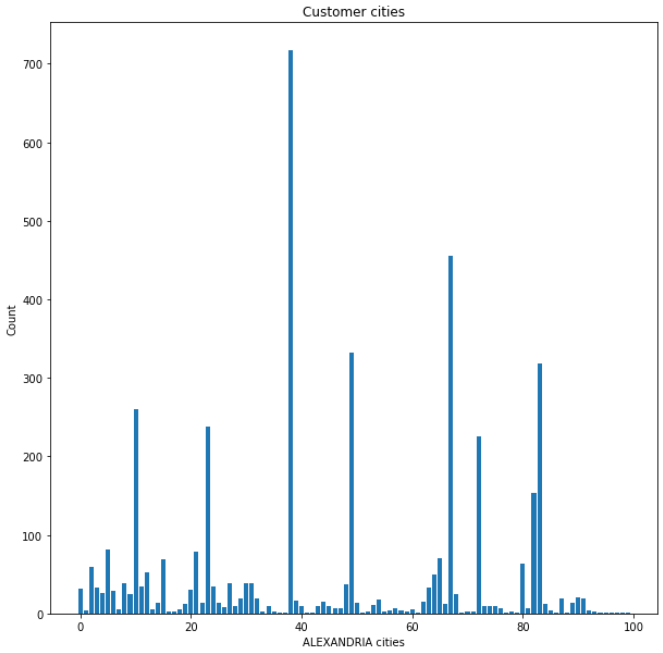
- **Expectation**: We expect that the collection of doctors contains country, province and city of them.
- **Collecting info**: We find that, database contains 67193 doctors with their locations.
- **Comparing Data and expectations**: Since the database contains doctors locations, we are able to answer the question.

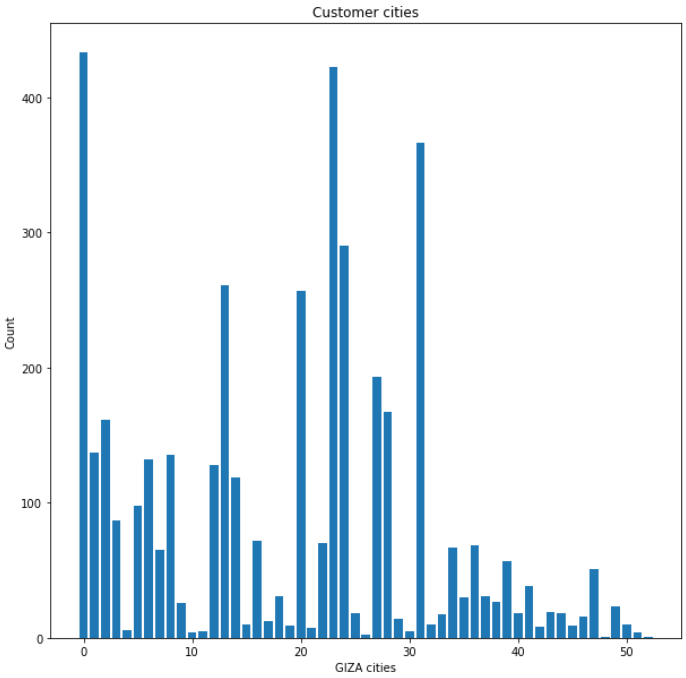
# 3- Model

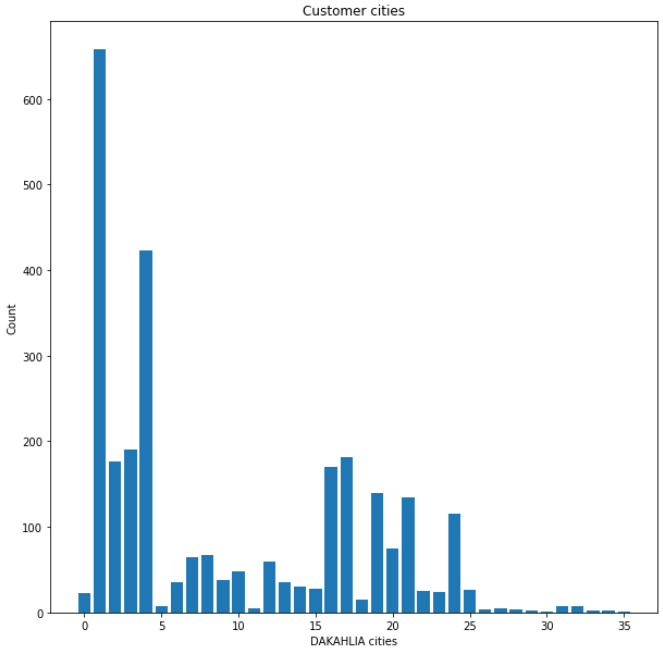
- **Expectation**: Build visualisation that answers our question
- **Collecting info**: We have the info we need.
- **Comparing Data and expectations**: Visualisation helped us answer our questions to get distribution of doctor's locations

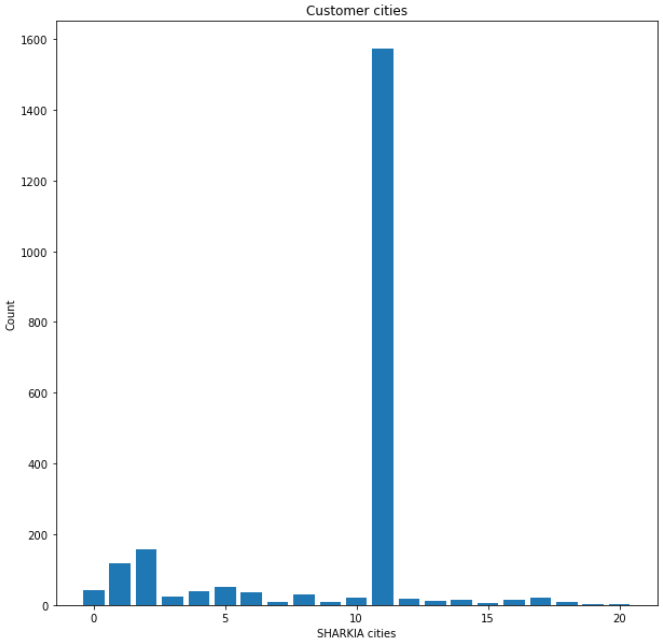





## 4- Interpreting results

- **Expectation**: For the province's distribution, we expect that Cairo contains the highest number of doctors.
- **Collecting Info**: The result visualisations
- **Comparing Data and expectations**: Visualisations results match the expectations.

## 5- Communicating results

- **Expectation**: Clear communication that leads to show doctor's distribution.
- **Comparing Data and expectations**: Our visualisations lead to an increased number of medical reps in populated places to cover whole doctors.

# Q5) Compare the distribution of doctors' locations in Egypt and whether the visits covered them well

## 1- Stating and Refining the question

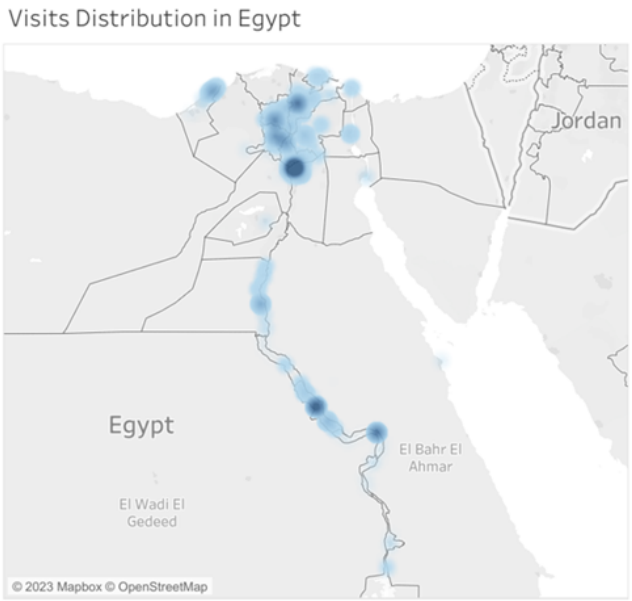
Same as Q3, as we have enough data for both doctors and visits to check this

## 2- Explore Data

Same as Q3 but we know already that we only target Egypt in this question.

## 3- Model

- **Expectation**: Build visualisation on Tableau that answers our question
- **Collecting info**: We have the info we need.
- **Comparing Data and expectations**: Visualisation helped us answer our questions to check locations not visited enough.

## 4- Interpreting results

- **Expectation**: Clear interpretation that tells us locations not covered if any.
- **Info:** Comparing the two images we can find that most of the regions are fairly covered however we can notice that some regions are not covered which are: Beheira, Kafr ElSheikh, Matrouh, Hurghada and some parts of Giza as well.
- **Comparing Data and expectations**: Our interpretation is clear and precise.

## 5- Communicating results

- **Expectation**: The interpretation is clear to make decisions.
- **Collecting info**: We have regions not well covered.
- **Comparing Data and expectations**: The results are clear and suggests that reps should increase in the regions mentioned above.

Tableau Dashboard for Questions 3 and 5

# Q6) Is there a relation between the number of doctors assigned and deviation of locations?

## 1- Stating and Refining the question

- **Expectation**: This question will help us to know more about the deviation of locations.
- **Collecting info**: We check the importance of the question and ability to answer it.
- **Comparing Data and expectations**: Our expectations match as the question will be very useful for pharma companies.

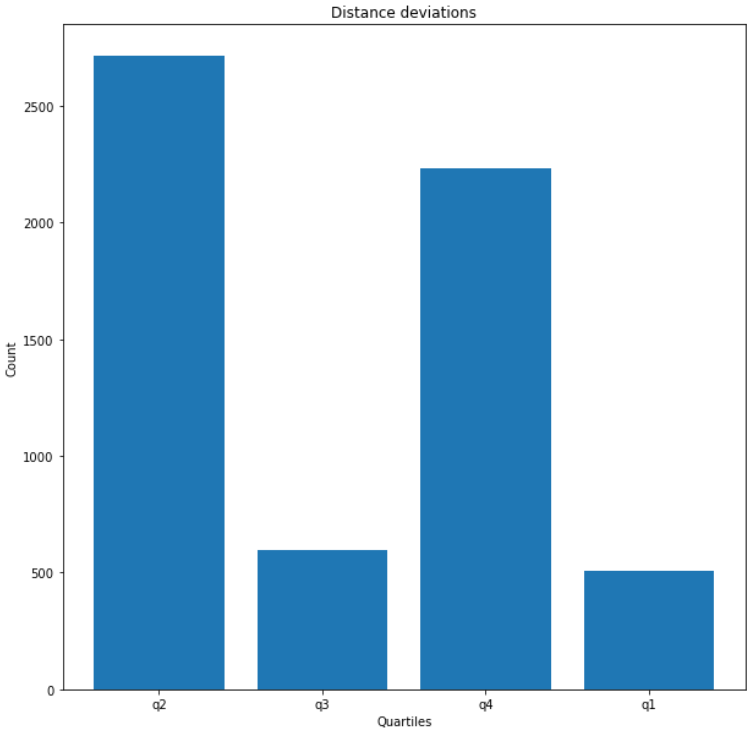
## 2- Explore Data

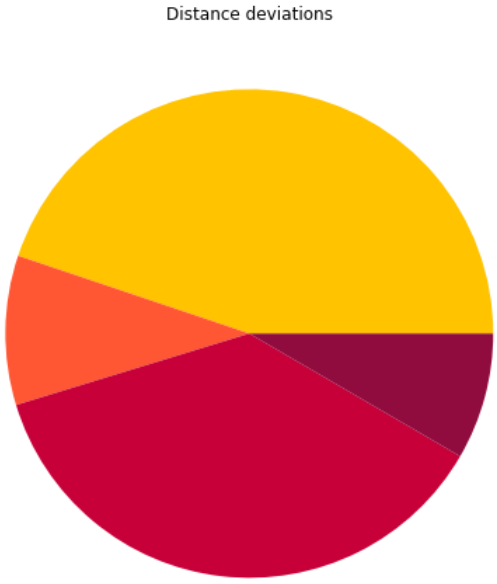
- **Expectation**: We expect that the collection of visits contains deviation of locations.
- **Collecting info**: We find that deviations of locations are stored in the trackers collection.
- **Comparing Data and expectations**: Since the database contains deviation of the visit, we are able to answer the question.

## 3- Model

- **Expectation**: Build visualisation that answers our question

- **Collecting info**: We have the info we need. We split the medical reps into four categories & each one assigned to (138, 155, 184, 215) doctors.
- **Comparing Data and expectations**: Visualisation helped us answer our questions to get the relation between number of doctors and distance deviation.




## 4- Interpreting results

- **Expectation**: We expect that there are such other factors that affect visit deviation.
- **Collecting Info**: The result visualisations
- **Comparing Data and expectations**: Visualisations results match the expectations.

## 5- Communicating results

- **Expectation**: Clear communication that leads to showing there is no clear relation between number of doctors and visits deviation.
- **Comparing Data and expectations**: Our visualisations lead to search more about the reason for visit deviation.

## Q7) Does time of visits affect productivity?

### 1- Stating and Refining the question

- **Expectation**: The question is if importance to the medical reps to choose the best working time for their productivity
- **Comparing Data and expectations**: The answer will be important but the question can be better refined to the following: What is the best working time for a medical rep?

### 2- Explore Data

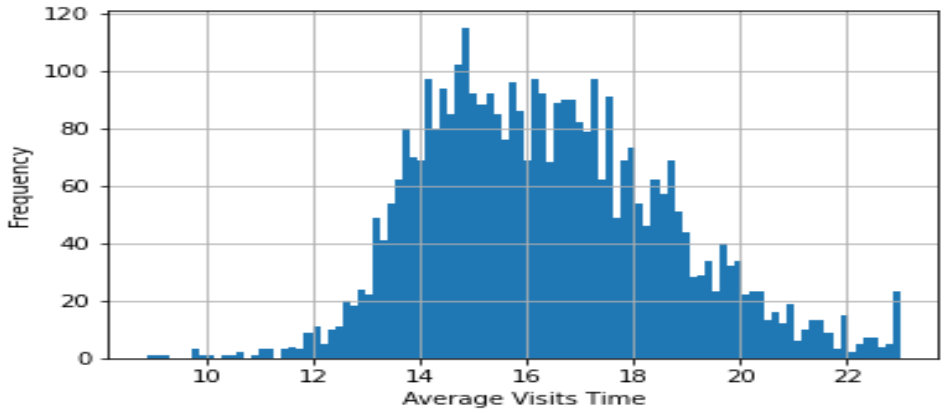
- **Expectation**: We expect to find enough data of visits where the exact time of visits are recorded and to find a measure of productivity
- **Collecting info**: We run a query on the database to retrieve the visits of a given company where visits are recorded and get the number of visits in each day in addition to total presentation time on that day.

- **Comparing Data and expectations**: Data will be good enough for the analysis

| | userid | day | totalDurationTime | averageVisitsTime | visitsNum |
|---|---|---|---|---|---|
| 0 | 621dd65e26d90d2838e4fdbd | 3/13/2023 | 2783 | 14.583333 | 12 |
| 1 | 6061c07af7606271c5c4da3d | 10/11/2022 | 5241 | 16.941176 | 17 |
| 2 | 628b3d307ee2c0232015abd5 | 10/12/2022 | 1631 | 13.090909 | 11 |
| 3 | 5da46ce7565e290015188597 | 4/1/2023 | 2140 | 14.083333 | 12 |
| 4 | 5da46f92565e2900151885b8 | 3/7/2023 | 1694 | 18.700000 | 10 |

### 3- Model

- **Expectation**: We need something to determine best working time for reps
- **Collecting info**: We will test some hypothesis using Z-Test on the above data, we can see on this histogram that most common average working time is around 2-4 pm



We will run the Z-Test since we have all dataset of visits for a given company on two hypothesis:
First Hypothesis is working around rush hour decreases productivity
H0: People who work around rush hour don't perform worse
H1: People who work around rush do have less productivity

By using the Z-Test (as we have the whole population which is all visits of that company) we retain the null hypothesis as the average of productivity using both measures mentioned above (total time of work and number of visits) isn't less the mean of productivity in other parts of the day (it is actually higher).

Second Hypothesis: People working at day shift perform better than working on night shift
H0: People who work on day shift perform same as night shift
H1: People who work on day shift perform better than night shift

Using Z-Test on first measure (Number of visits), we get:
Z-Score = 4.77
P-value = 9.10e-07 << 0.05

So we reject the null hypothesis.

Using Z-Test on second measure (Total presentation time), we get:
Z-Score = 3.2766
P-value = 0.0005 < 0.05

So we reject the null hypothesis as well and conclude that this is an interesting hypothesis and there might be a clear relation between time of working and productivity however it could be due to multiple factors not related to the time itself like the number of doctors working at day compared to at night for example.

- **Comparing Data and expectations**: Results gave us a hint about best working time given the available data.

## 4- Interpreting results

- **Expectation**: Results gives a hint on what working time increases productivity
- **Info:** Our interpretation is that working in the morning might be better in general productivity wise.
- **Comparing Data and expectations**: Expectations match

## 5- Communicating results

- **Expectation**: The interpretation is clear to make decisions.
- **Collecting info**: We have a suggestion but not without a formal prove that working in the morning in better
- **Comparing Data and expectations**: The results will be important but might not be enough to make decisions because there are other factors not available in our data that can affect this interpretation such as Doctors working time preferences, also rejecting null hypothesis doesn't necessarily confirm the alternative hypothesis.

# Q8) Does the supervisor have an effect on rep performance?

## 1- Stating and Refining the question

- **Expectation**: This question will help us to know the importance of supervisors & how they affect medical reps' performance.
- **Collecting info**: We check the importance of the question and ability to answer it.
- **Comparing Data and expectations**: Our expectations match as the question will be very useful for pharma companies.

## 2- Explore Data

- **Expectation**: We expect that the collection of visits and companies will help us to measure rep's performance & relate it to the supervisor.
- **Collecting info**: We find that collections contain the data that we need.
- **Comparing Data and expectations**: Since the database contains the data, we are able to answer the question.

## 3- Model

- **Expectation**: Build a dataset that contains the following fields ["Rep", "Supervisor", "Month", "Customers", "Target_Frequency", "Visits", "Coverage"] to get the relation between the rep's performance & supervisor.
- **Collecting info**: From visits and companies collections, we built our dataset.
- **Comparing Data and expectations**: Using Chi-Squared Statistics, we found that there is a dependency between coverage column & supervisor column. Chi-Squared Statistic: 33.70972743202966, p-value: 4.58524965488053e-05. In general, if the p-value is less than the chosen significance level (often 0.05), then we reject the null hypothesis and conclude that there is evidence for an association between the two variables.

### 4- Interpreting results

- **Expectation**: We expect that there is a dependency between rep & supervisor.
- **Collecting Info**: Model result
- **Comparing Data and expectations**: Results match the expectations.

### 5- Communicating results

- **Expectation**: Clear communication that leads to showing the effect of the supervisor on medical reps & how to increase not just medical rep's performance but also supervisor's performance.
- **Comparing Data and expectations**: Our results lead to search for how to increase supervisor's performance.

## Q9) How to increase performance by reassigning customers to reps?

### 1- Stating and Refining the question

- **Expectation**: This question will help us to know medical reps' performance and the relation between the covered area & rep performance.
- **Collecting info**: We check the importance of the question and ability to answer it.
- **Comparing Data and expectations**: Due to the given database, we are able to predict the medical reps' performance based on coverage area.

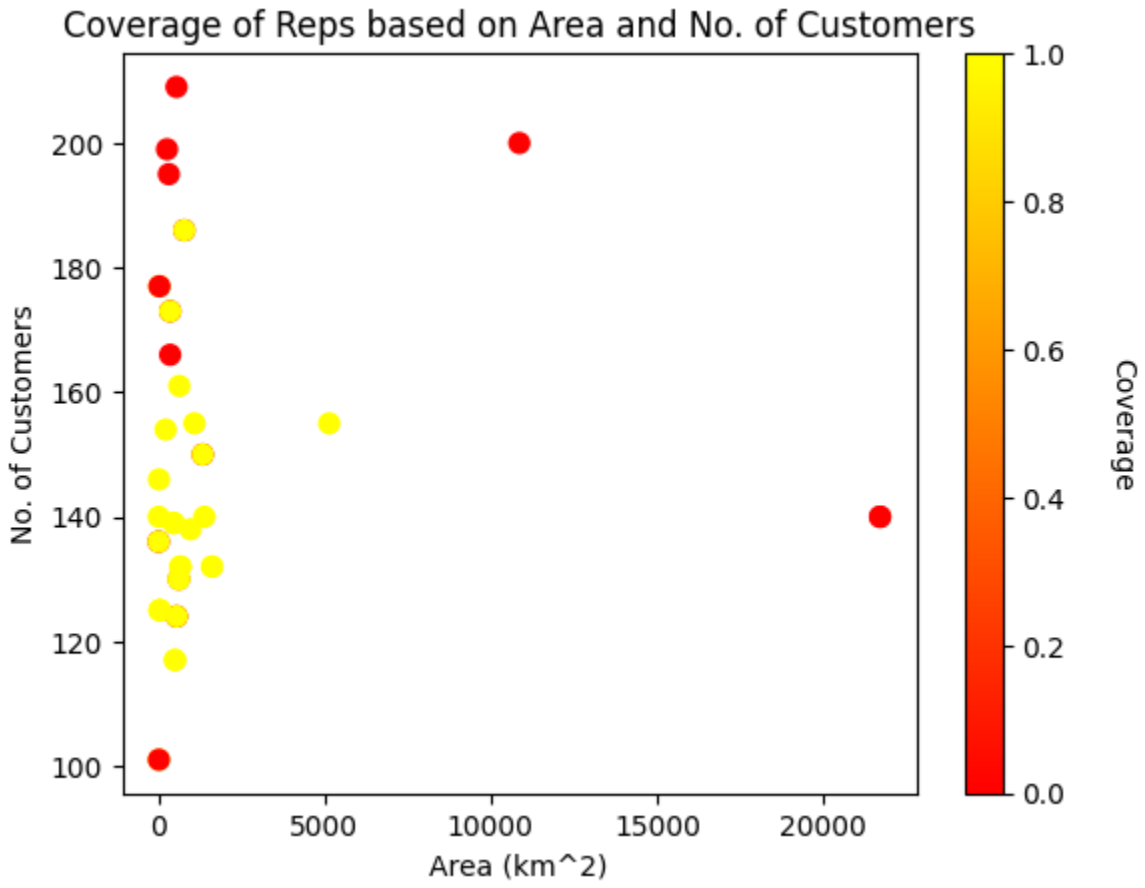
### 2- Explore Data

- **Expectation**: We expect that the collection of visits, customers and companies will help us to measure rep's performance & relate it to the coverage area.
- **Collecting info**: We find that collections contain about 3654 doctors with their geoLocation.
- **Comparing Data and expectations**: Since the database contains the data, we are able to answer the question.

### 3- Model

- **Expectation**: Build a dataset that contains the following fields ["Rep", "Supervisor", "Month", "Customers","Area" ,"Target_Frequency", "Visits", "Coverage"] to predict coverage from the rep's performance.
- **Collecting info**: From visits, customer and companies collections, we built our dataset.

- **Comparing Data and expectations**: Using SVM, we build our model & get 100% accuracy on testset. Visualisation helped us answer our questions to get the relation rep's performance & area coverage.




## 4- Interpreting results

- **Expectation**: We expect that increasing coverage area will lead to zero coverage.
- **Collecting Info**: Model result
- **Comparing Data and expectations**: Results match the expectations.

## 5- Communicating results

- **Expectation**: Clear communication that leads to showing the effect of the coverage area on medical reps & how to decrease it to increase medical rep's performance.
- **Comparing Data and expectations**: Our results lead to search for how to decrease coverage area.

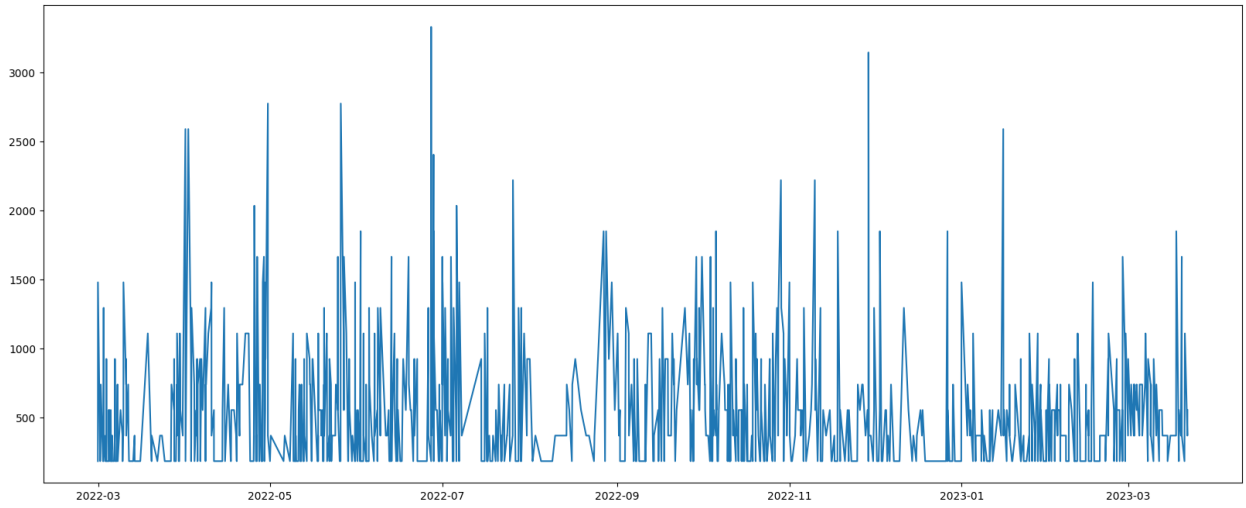
## Q10) Predict the products total profit throughout the time?

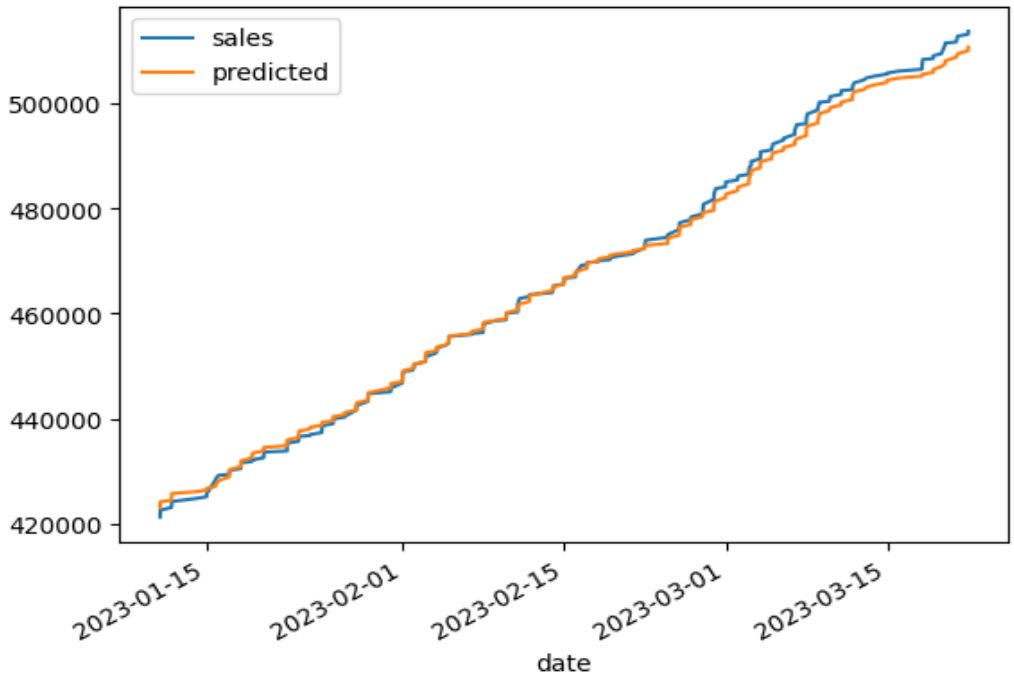
# 1- Stating and Refining the question

- **Expectation**: We want to expect the product total profits thought the time
  So, may it give us a good indication about our performance
- **Collecting info**: We collect the top 5 product in terms of sales and analyze accumulate their total prices, then make a prediction
- **Comparing Data and expectations**: Due to the given database, we are able to predict the our monthly profits with minimum error.

# 2- Explore Data

- **Expectation**: We expect that the product sales will have some fluctuations due to the different needs through different time a long the year
- **Collecting info**: collect the total profits from each sale for the product, along with the sailing date.
- **Comparing Data and expectations**: the products like expected varies in total profits throughout the year time.






## 3- Model

- **Expectation**: as stated before the nature of the problem is time series forecasting, so build a Model using LSTM with 64 node, followed by relu activation function and one linear function for final prediction
- **Collecting info**: use the data collected from the previous step as dataframe from the date and accumulated profits.
- **Comparing Data and expectations**: as shown in the  above figure the actual sales is too close to the predicted one which stated that our model behaves well, and may give us a precise prediction about the future.

## 4- Interpreting results

- **Expectation**: We expect that total accumulated profits will increase with constant steps.
- **Collecting Info**: Model result
- **Comparing Data and expectations**: Results match the expectations.

## 5- Communicating results

- **Expectation**: Clear communication that leads to show that there will be a good and profits due to our sales history
- **Comparing Data and expectations**: It seems we’re doing well.

## Future work and enhancements:

- Work and collect more data
- If we can get more financial data we can do much more.
- Try to attract more customers
- Build an application for data analysis