Northumbria University NEWCASTLE

# COURSEWORK ASSESSMENT SPECIFICATION

| | |
|---|---|
| Student Number: | 21062830 |
| Student Name: | Anika Tanzim |
| Module Title: | Big Data Analytics |
| Module Number: | LD7186 |
| Module Tutor Name(s): | Nitsa Herzog, Rose Fong |
| Academic Year: | 2021-2022 |
| Programme title: | Msc Big Data and Data Science Technology with AP |
| Coursework Title: | Assignment |
| Submission Deadline: | 12th May,2022, 16:00 |
| Actual date of submission: | 12th May,2022 |

# Table of Contents

# Section 1 Big Data Analytics

## 1.1 Task-1: Problem Domain, Data Description, and Research Question

### 1.1.1 Problem Domain

The obtained dataset contains information about employees of a specific company. The application domain of the dataset is **Human Resources(HR) Analytics** for the particular company or organization.

HR analytics is a data-driven approach of HR Management which aids to measure the HR metrics to business performance and guides to take a good decision towards business. HR analytics is the process of deliberately measuring the human factors that influence corporate goals (Van Den Heuvel and Bondarouk, 2016).


*Figure 1:HR Analytics- The Application Domain Source: (Gupta, 2022)*

### 1.1.2 Data Description

This HR Analytics dataset (HR Analytics, 2022) is a publicly available dataset on Kaggle (Kaggle: Your Machine Learning and Data Science Community, 2022). This is a **secondary** structured data in a **standard** format with 10 variables uploaded by Giri Pujar in 2018. It is a csv file named “HR_comma_sep.csv” of 566.79kB file size. No information is available to refer the dataset as real or anonymised. This dataset is designed to know the factors that lead an employee to leave the company.

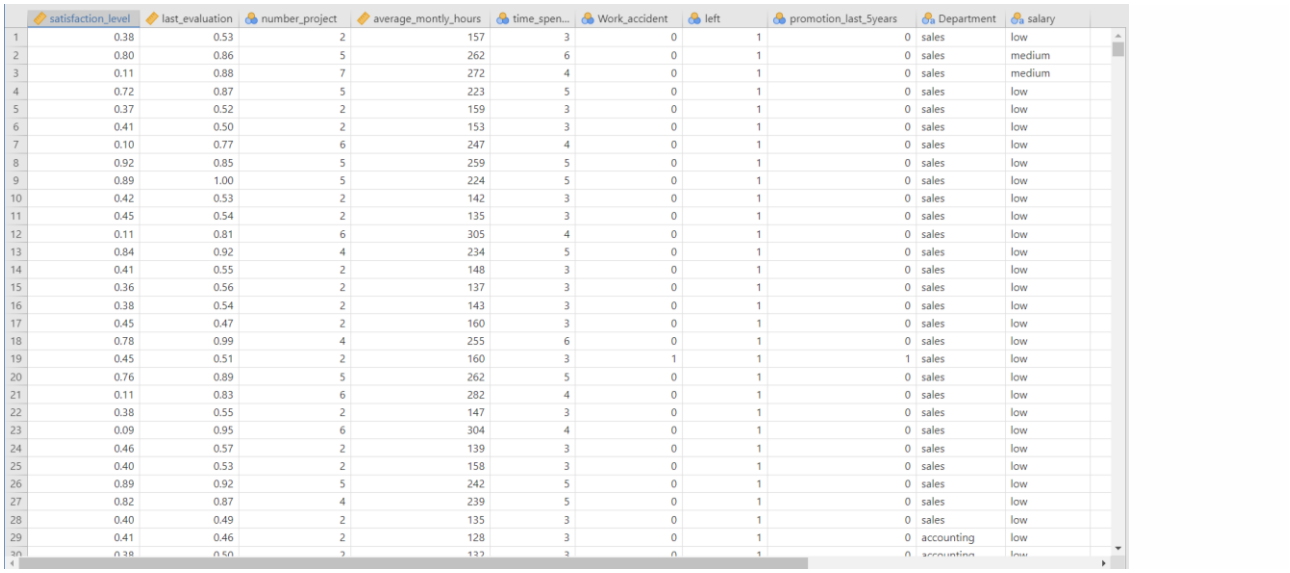
| Variable Name | Description |
|---|---|
| **Satisfaction_level** | Satisfaction |

| | |
|---|---|
| **Last_evaluation** | Evaluation percentage |
| **Number_project** | Number of project handled |
| **Average_monthly_hours** | Monthly average hours |
| **Time_spend_company** | Number of years worked |
| **Work_accident** | Work accident |
| **Left** | Turnover |
| **Promotion_last_5years** | Promotion in last 5 years |
| **Department** | Department |
| **Salary** | Salary |

The exploratory data analysis of this dataset is done with **Python** and **Jamovi**.


*Figure 2: Loading Dataset in Python*

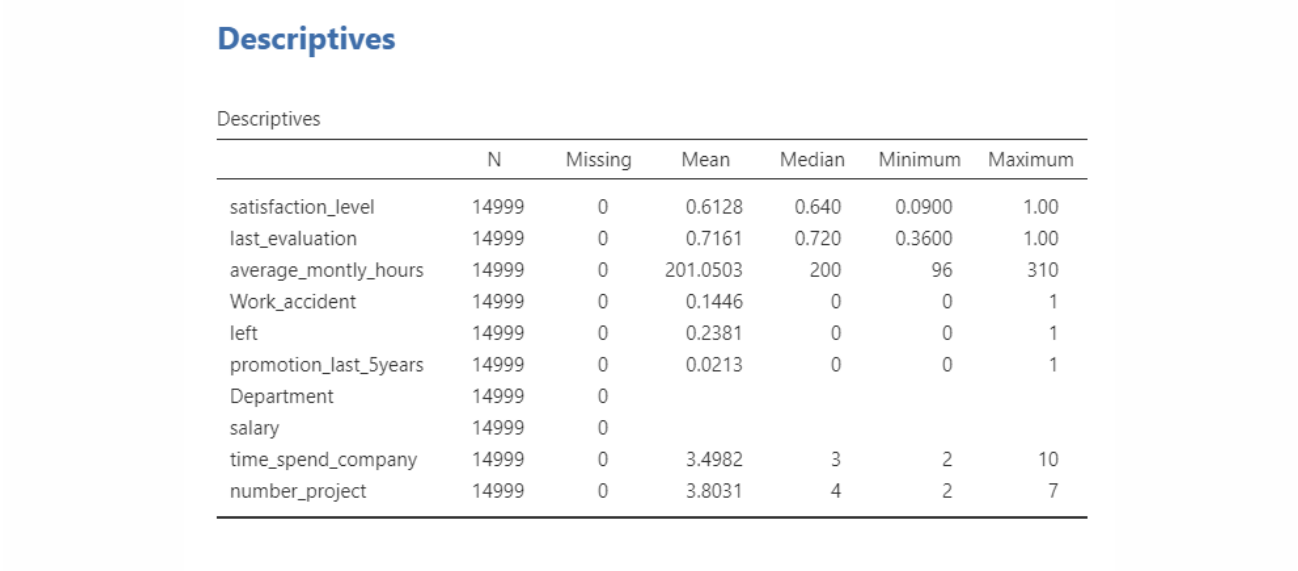
**Google Colab** notebook is used to use **Python** script. Dataset is stored in the google drive and then it is loaded by mounting the drive.



| | satisfaction_level | last_evaluation | number_project | average_montly_hours | time_spen... | Work_accident | left | promotion_last_5years | Department | salary |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.38 | 0.53 | 2 | 157 | 3 | 0 | 1 | 0 | sales | low |
| 2 | 0.80 | 0.86 | 5 | 262 | 6 | 0 | 1 | 0 | sales | medium |
| 3 | 0.11 | 0.88 | 7 | 272 | 4 | 0 | 1 | 0 | sales | medium |
| 4 | 0.72 | 0.87 | 5 | 223 | 5 | 0 | 1 | 0 | sales | low |
| 5 | 0.37 | 0.52 | 2 | 159 | 3 | 0 | 1 | 0 | sales | low |
| 6 | 0.41 | 0.50 | 2 | 153 | 3 | 0 | 1 | 0 | sales | low |
| 7 | 0.10 | 0.77 | 6 | 247 | 4 | 0 | 1 | 0 | sales | low |
| 8 | 0.92 | 0.85 | 5 | 259 | 5 | 0 | 1 | 0 | sales | low |
| 9 | 0.89 | 1.00 | 5 | 224 | 5 | 0 | 1 | 0 | sales | low |
| 10 | 0.42 | 0.53 | 2 | 142 | 3 | 0 | 1 | 0 | sales | low |
| 11 | 0.45 | 0.54 | 2 | 135 | 3 | 0 | 1 | 0 | sales | low |
| 12 | 0.11 | 0.81 | 6 | 305 | 4 | 0 | 1 | 0 | sales | low |
| 13 | 0.84 | 0.92 | 4 | 234 | 5 | 0 | 1 | 0 | sales | low |
| 14 | 0.41 | 0.55 | 2 | 148 | 3 | 0 | 1 | 0 | sales | low |
| 15 | 0.36 | 0.56 | 2 | 137 | 3 | 0 | 1 | 0 | sales | low |
| 16 | 0.38 | 0.54 | 2 | 143 | 3 | 0 | 1 | 0 | sales | low |
| 17 | 0.45 | 0.47 | 2 | 160 | 3 | 0 | 1 | 0 | sales | low |
| 18 | 0.78 | 0.99 | 4 | 255 | 6 | 0 | 1 | 0 | sales | low |
| 19 | 0.45 | 0.51 | 2 | 160 | 3 | 1 | 1 | 1 | sales | low |
| 20 | 0.76 | 0.89 | 5 | 262 | 5 | 0 | 1 | 0 | sales | low |
| 21 | 0.11 | 0.83 | 6 | 282 | 4 | 0 | 1 | 0 | sales | low |
| 22 | 0.38 | 0.55 | 2 | 147 | 3 | 0 | 1 | 0 | sales | low |
| 23 | 0.09 | 0.95 | 6 | 304 | 4 | 0 | 1 | 0 | sales | low |
| 24 | 0.46 | 0.57 | 2 | 139 | 3 | 0 | 1 | 0 | sales | low |
| 25 | 0.40 | 0.53 | 2 | 158 | 3 | 0 | 1 | 0 | sales | low |
| 26 | 0.89 | 0.92 | 5 | 242 | 5 | 0 | 1 | 0 | sales | low |
| 27 | 0.82 | 0.87 | 4 | 239 | 5 | 0 | 1 | 0 | sales | low |
| 28 | 0.40 | 0.49 | 2 | 135 | 3 | 0 | 1 | 0 | sales | low |
| 29 | 0.41 | 0.46 | 2 | 128 | 3 | 0 | 1 | 0 | accounting | low |

*Figure 3: Data View of the dataset in Jamovi*

Also the dataset is also loaded in **Jamovi** software by importing the csv file. 2 variables are string type and other are numeric. 3 of them are Scale and others are Nominal.



**Descriptives**

Descriptives

| | N | Missing | Mean | Median | Minimum | Maximum |
|---|---|---|---|---|---|---|
| satisfaction_level | 14999 | 0 | 0.6128 | 0.640 | 0.0900 | 1.00 |
| last_evaluation | 14999 | 0 | 0.7161 | 0.720 | 0.3600 | 1.00 |
| average_montly_hours | 14999 | 0 | 201.0503 | 200 | 96 | 310 |
| Work_accident | 14999 | 0 | 0.1446 | 0 | 0 | 1 |
| left | 14999 | 0 | 0.2381 | 0 | 0 | 1 |
| promotion_last_5years | 14999 | 0 | 0.0213 | 0 | 0 | 1 |
| Department | 14999 | 0 | | | | |
| salary | 14999 | 0 | | | | |
| time_spend_company | 14999 | 0 | 3.4982 | 3 | 2 | 10 |
| number_project | 14999 | 0 | 3.8031 | 4 | 2 | 7 |

*Figure 4: Characteristics of each features in the dataset*

This dataset contains 14999 entries. The average of the employee satisfaction is around 61%. Around 3 or 4 projects are done by each employee on an average. Around 3.5 years are spent by most of the employees before leaving the company. Lastly, around 23.8% of the total employee has left the company for a particular reason.

### 1.1.3 Research Question & Hypothesis

**Frequencies**

Frequencies of left

| Levels | Counts | % of Total | Cumulative % |
|---|---|---|---|
| 0 | 11428 | 76.2 % | 76.2 % |
| 1 | 3571 | 23.8 % | 100.0 % |

*Figure 5: Frequencies of left*

According to the frequencies of "left", the turnover rate of this company is 23.8 ~ 24%. So, it is obliged to make the turnover rate less for this compnay. Therefore, the question arises "**Why are the employees leaving?**". The goal of this anaysis is to find the factors or the variables that are responsible for employee turnover.

**RQ#1** Is "salary" the factor for making the employee leave the company?

**$H_0$:** The factor salary is not the reason for an employee to leave the company.

**$H_1$:** An employee can leave because of the low salary.

**RQ#2** Is "satisfaction" the most important atrribute for making the employee leave the company?

**$H_0$:** The factor "satisfaction" is not the reason for an employee to leave the company.

**$H_1$:** An employee can leave because he/she is not satisfied with his/her job.

**RQ#3** Do number of project and evalution impact together on turnover of the company?

**$H_0$:** Number of project and evalution do not impact together on turnover of the company.

**$H_1$:** Number of project and evalution impact together on turnover of the company.

## 1.2 Task-2: Solution Exploration

### 1.2.1 Approaches and Technologies of Big Data Applications

The 5V's define Big Data: volume, velocity, value, veracity, variety. Big Data are created rapidly and massive. It is important to generate the Big data lifecycle from data generation to data visualization through some approaches. There are some traditional data analysis approaches like- **Cluster Analysis, factor Analysis, Correlation Analysis, Regression Analysis, A/B Testing, Statistical Analysis, and Data Mining**. However, traditional approaches had some limitations because of which "**Big Data Analytics**" was introduced. There are three types of analytics techniques like- **Descriptive, Predictive and Prescriptive Analysis** (Rabhi, Falih, Afraites and Bouikhalene, 2019)**.**

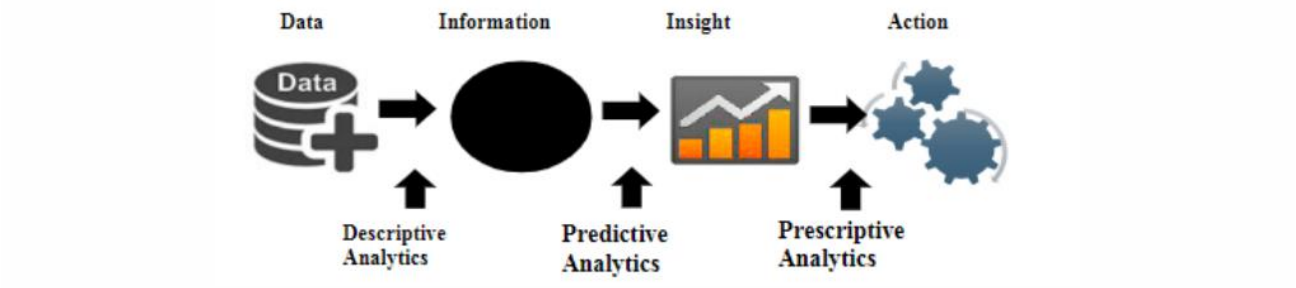



*Figure 6: Types of Analytics Techniques Source: (Rabhi, Falih, Afraites and Bouikhalene, 2019)*

Big data technologies are classified into 4 domains *(Technologies, 2022)*. Such as – Data Storage, Analytics, Data mining, Visualization. Under the data storage domain, Hadoop, MongoDB, Hunk, and Cassandra are top technologies that has been used. For the Data Mining field, Presto, ElasticSearch, Apache Kafka, Splunk, Apache Spark and R language are one the best environments. Data visualization related technologies are – Tableau, Plotly, SPSS etc.

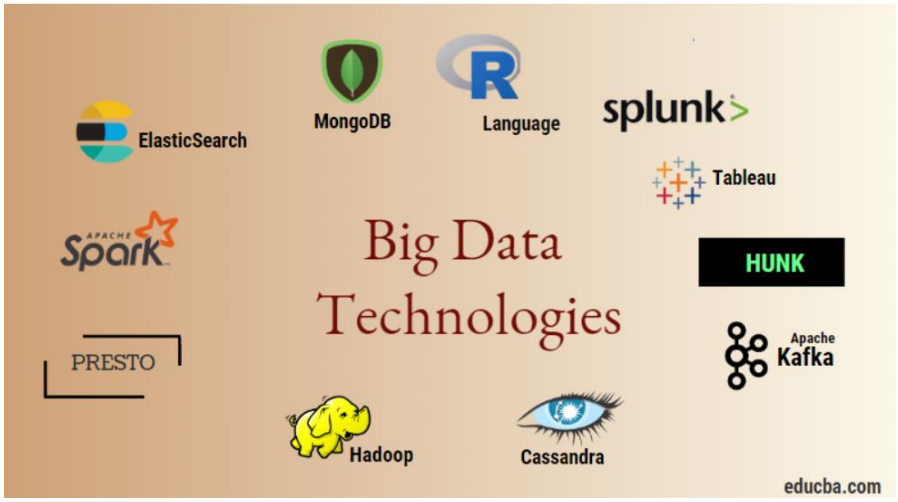



*Figure 7: Big Data Technologies Source: (Technologies, 2022)*

### 1.2.2 Solutions and Techniques Regarding the Obtained Dataset

HR Analytics is the top most concept in the socio-economic word. A case study proposed a smart framework by using **JASP** where the margin error of traditional analysis was reduced (Jabir, Falih and Rahmani, 2019). This case study defines the HR analytical power and its benefits for the company using **paired sample t-test**.

Another paper made a literature review on HR Analytics and proposed a new model where **SPSS** was used for analysis (Opatha, 2020). **Correlation Matrix** and **ANOVA test** were done for implementing significance.

Employee **attrition** analysis was done using **logistic regression** in a paper where a prediction model was proposed (Setiawan, Suprihanto, Nugraha and Hutahaean, 2020). In this paper 11 variables were highlighted as important factor in employee attrition.

A paper shows the impacts of turnover, retention and business process over HR analytics with the help of **Decision Tree Classifier** and **KNN algorithm** (Roshini et al., 2021). It says that there is a significant correlation between the chosen and independent variable by using data mining technique.

### 1.2.3 Taken Methodological Approach

The approach taken for this HR Analytics Problem is **Statistics**. This technique will help to organize, experiment and interpret data. It includes regression analysis, factor analysis, clustering, recognition analysis etc. **Descriptive statistics** will be used for visualization in two ways- graphically and numerically.

To describe qualitative data **Bar graph** and **Pie Chart** will be used, and to describe quantitative data **Histogram** will be used. For grouping numerical and categorical data **Frequency distribution** will be used.

Hypothesis testing will be done for statistical significance testing with **T-test or Chi-square test or ANOVA test** depending on the research question.

**Box-plot, Scatter-plot, Correlation matrix, Probability-plot** and **Kernel Density** Plot will also be used for descriptive analysis.

The technologies that will be used for analysis and visualization are **Jamovi** and **Python**.

## 1.3 Task-3: Solution Development

### 1.3.1 General Data Analysis

**Satisfaction_level:**

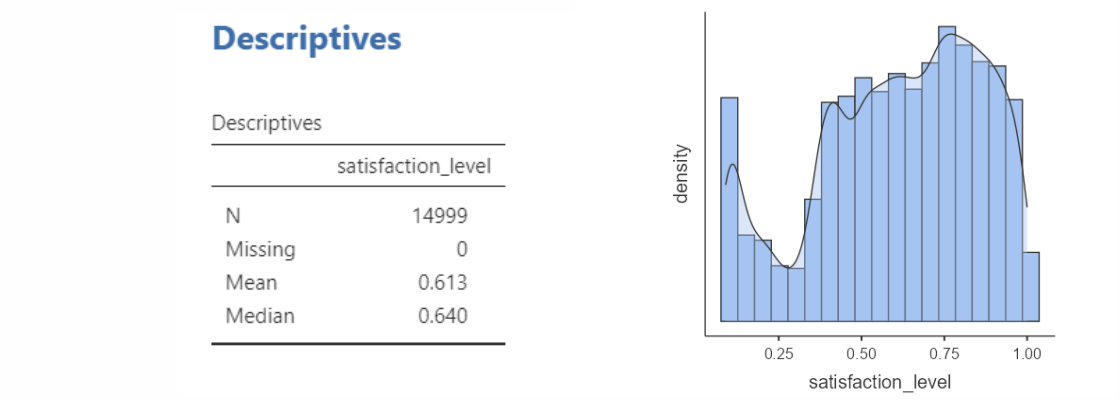



*Figure 8: Descriptive and Histogram of satisfaction_level*

The mean value of satisfaction level is 61.3%. Most of the employees are satisfied in between 35% to 85% which means it varies for individual stuff.

Descriptives

| | left | satisfaction_level |
|---|---|---|
| N | 0 | 11428 |
| | 1 | 3571 |

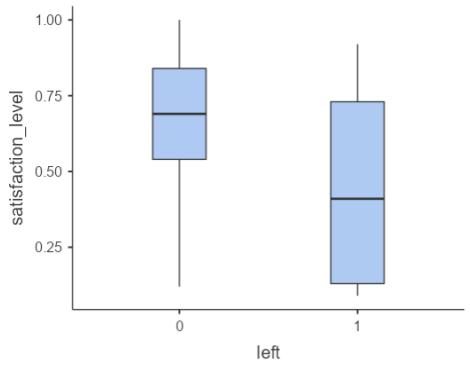



*Figure 9: Boxplot of satisfaction_level in terms of left*

To relate the satisfaction with the turnover, the boxplot is used. The persons who left the company had a consistent lower satisfaction.

```
[16] #KDEPlot: Kernel Density Estimate Plot
     fig = plt.figure(figsize=(15,4))
     ax=sns.kdeplot(df.loc[(df['left'] == 0),'satisfaction_level'] , color='b',shade=True, label='no turnover')
     ax=sns.kdeplot(df.loc[(df['left'] == 1),'satisfaction_level'] , color='r',shade=True, label='turnover')
     plt.title('Satisfaction - Turnover vs No Turnover')

     Text(0.5, 1.0, 'Satisfaction - Turnover vs No Turnover')
```

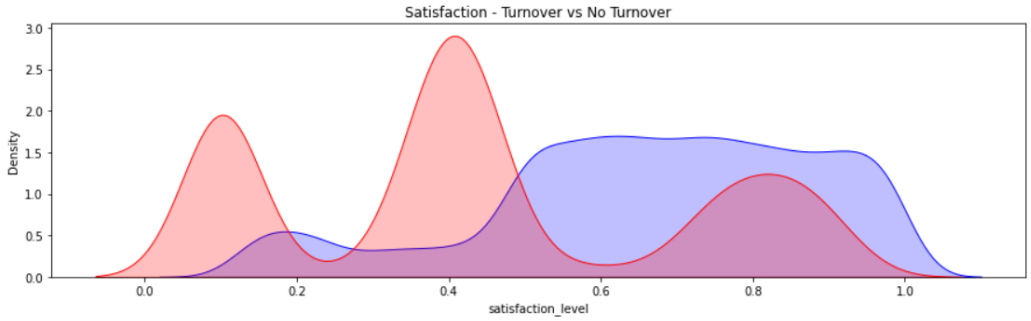



*Figure 10: Kernel density plot of employee satisfaction level - turnover vs no turnover*

It is a tri-model distribution where 0.0-0.2, 0.3-0.5 and 0.7-0.9 satisfaction level bearers tend to leave the company more.

**last_evaluation:**

Descriptives

| | last_evaluation |
|---|---|
| N | 14999 |
| Missing | 0 |
| Mean | 0.716 |
| Median | 0.720 |

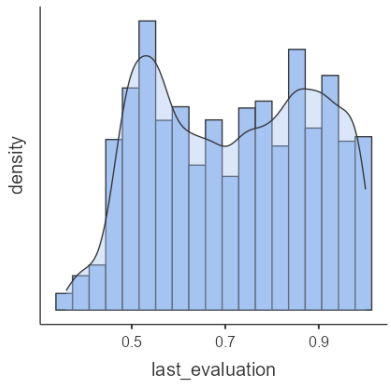



*Figure 11: Descriptive and Histogram of evaluation*

Employees are evaluated as 71% by the employers on an average. Several employees got evaluated around 50% or 80%.

```
# Kernel Density Plot
fig = plt.figure(figsize=(15,4),)
ax=sns.kdeplot(df.loc[(df['left'] == 0),'last_evaluation'] , color='b',shade=True,label='no turnover')
ax=sns.kdeplot(df.loc[(df['left'] == 1),'last_evaluation'] , color='r',shade=True, label='turnover')
plt.title('Evaluation - Turnover vs No Turnover')
```

```
Text(0.5, 1.0, 'Evaluation - Turnover vs No Turnover')
```

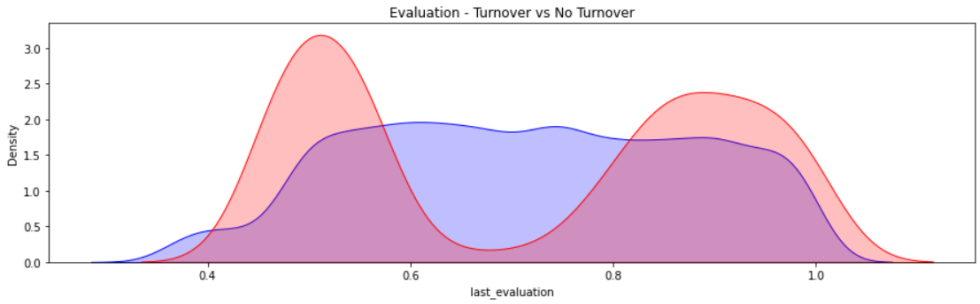



*Figure 12: Kernel density plot of Employee Evaluation - turnover vs no turn over*

The leaving employees follows a bimodal distribution since they tend to be evaluated in two distinct scale. The people who stayed are more likely to have 0.6-0.8 evaluation.

**number_project:**

**Descriptives**

Descriptives

| | number_project |
|---|---|
| N | 14999 |
| Missing | 0 |
| Mean | 3.80 |
| Median | 4 |

**Frequencies**

Frequencies of number_project

| Levels | Counts | % of Total | Cumulative % |
|---|---|---|---|
| 2 | 2388 | 15.9 % | 15.9 % |
| 3 | 4055 | 27.0 % | 43.0 % |
| 4 | 4365 | 29.1 % | 72.1 % |
| 5 | 2761 | 18.4 % | 90.5 % |
| 6 | 1174 | 7.8 % | 98.3 % |
| 7 | 256 | 1.7 % | 100.0 % |

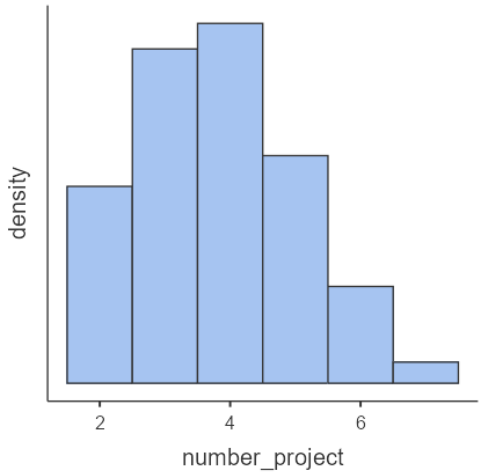



*Figure 13: Descriptive and Histogram of number_project*

Most of the employees handled 3 or 4 projects. Only a few handled more than 6 projects.

**average_monthly_hours:**

**Descriptives**

Descriptives

| | average_montly_hours |
|---|---|
| N | 14999 |
| Missing | 0 |
| Mean | 201 |
| Median | 200 |

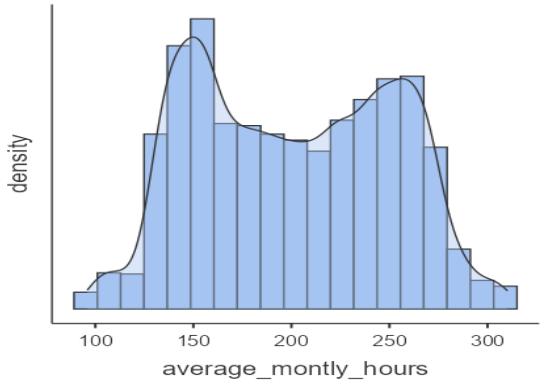



*Figure 14: Descriptive and Histogram of average_monthly_hours*

Employees spent about 150 to 250 hours monthly on an average in this company.

```
[14] #KDEPlot: Kernel Density Estimate Plot
     fig = plt.figure(figsize=(15,4))
     ax=sns.kdeplot(df.loc[(df['left'] == 0),'average_montly_hours'] , color='b',shade=True, label='no turnover')
     ax=sns.kdeplot(df.loc[(df['left'] == 1),'average_montly_hours'] , color='r',shade=True, label='turnover')
     plt.title('Average Monthly Hours  - Turnover vs No Turnover')
```

```
Text(0.5, 1.0, 'Average Monthly Hours  - Turnover vs No Turnover')
```

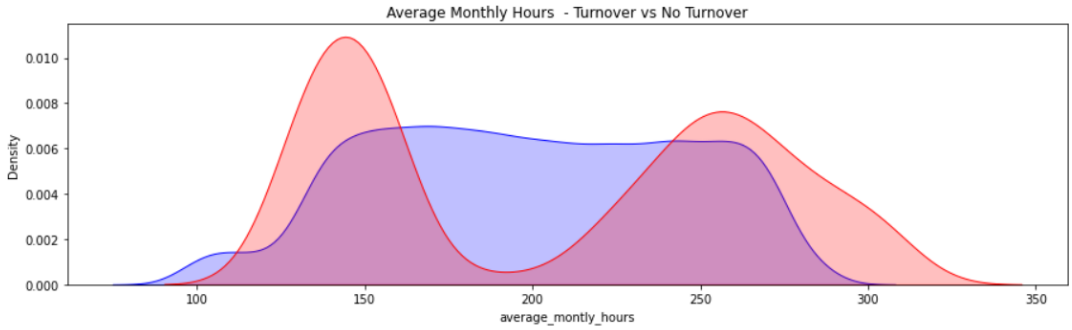



*Figure 15: Kernel density plot of Employee Average Monthly Hours - turnover vs no turn over*

People who left are more likely to be underworked or overworked. The employee who stayed tend to have 150 to 250 hours.

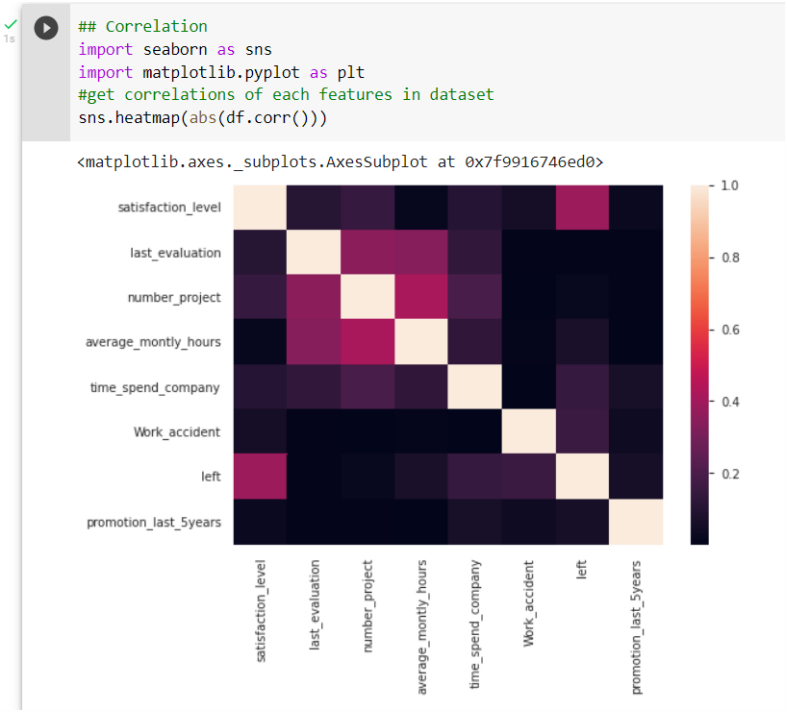



*Figure 16: Pearson Correlation of the dataset*

Pearson correlation matrix is visualized where number_project, last_evaluation and average_monthly_hours are correlated moderately among themselves.

Correlation Matrix

| | average_montly_hours | last_evaluation | number_project |
|---|---|---|---|
| average_montly_hours | — | | |
| last_evaluation | 0.340 | — | |
| number_project | 0.417 | 0.349 | — |

*Figure 17: Correlation Matrix*

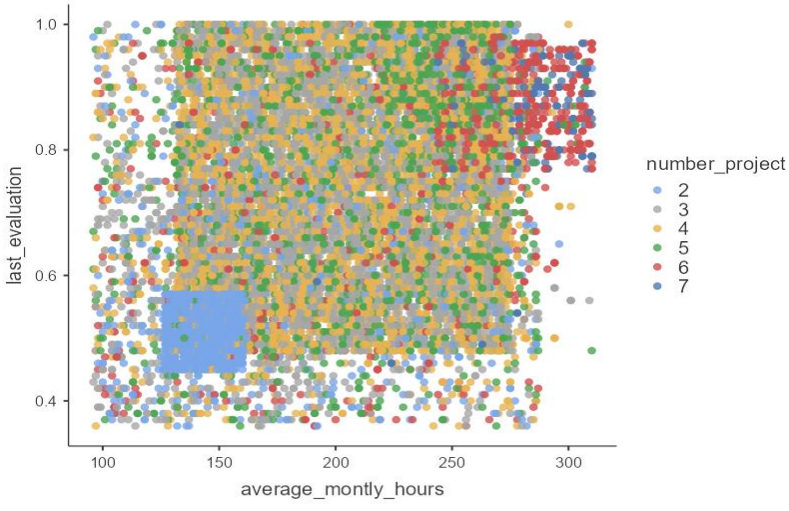



*Figure 4: Scatterplot of average_monthly_hours vs last_evaluation grouped by project count*

The scatterplot shows two significant clusters i.e.

- The blue one where average monthly hours is around150 and evaluation is centred on 0.5. The project count for this cluster is 2.
- The red one where average monthly hours are above 250 with high evaluation score which is centred on 0.9. The project count for this cluster is above 6.

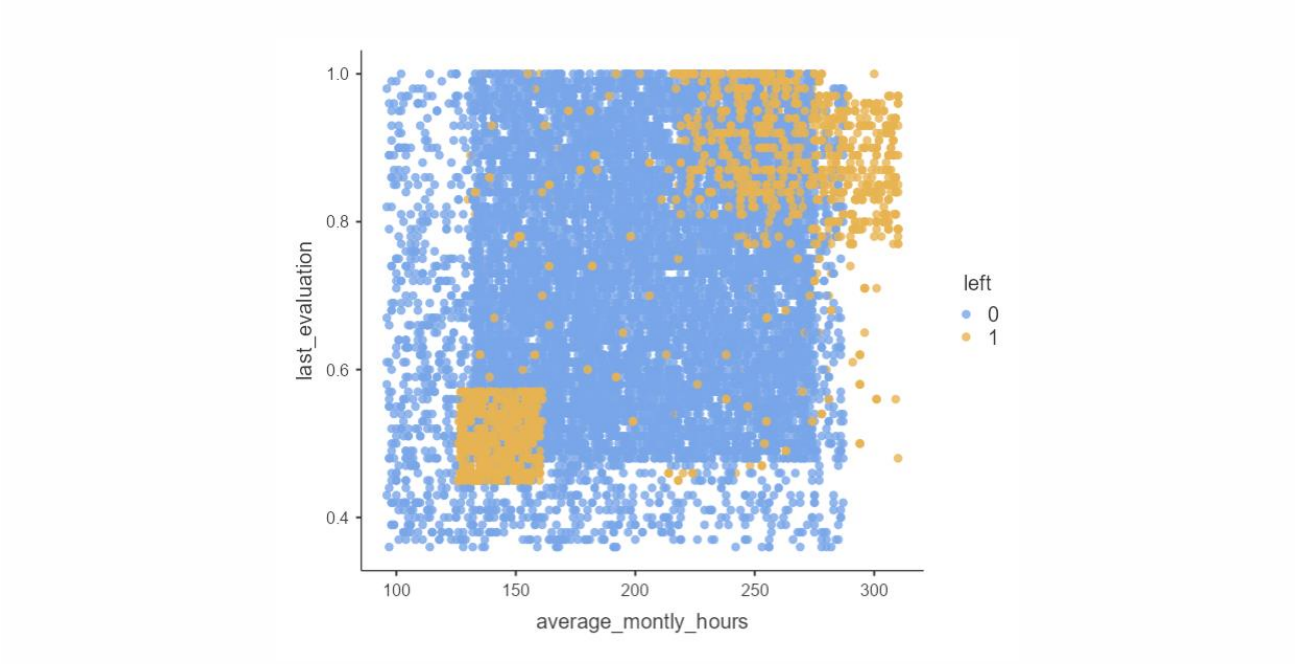



*Figure 5: Scatterplot of average_monthly_hours vs last_evaluation grouped by left*

This scatterplot shows as same clusters as the previous one which means the above three attributes are related to the turnover.

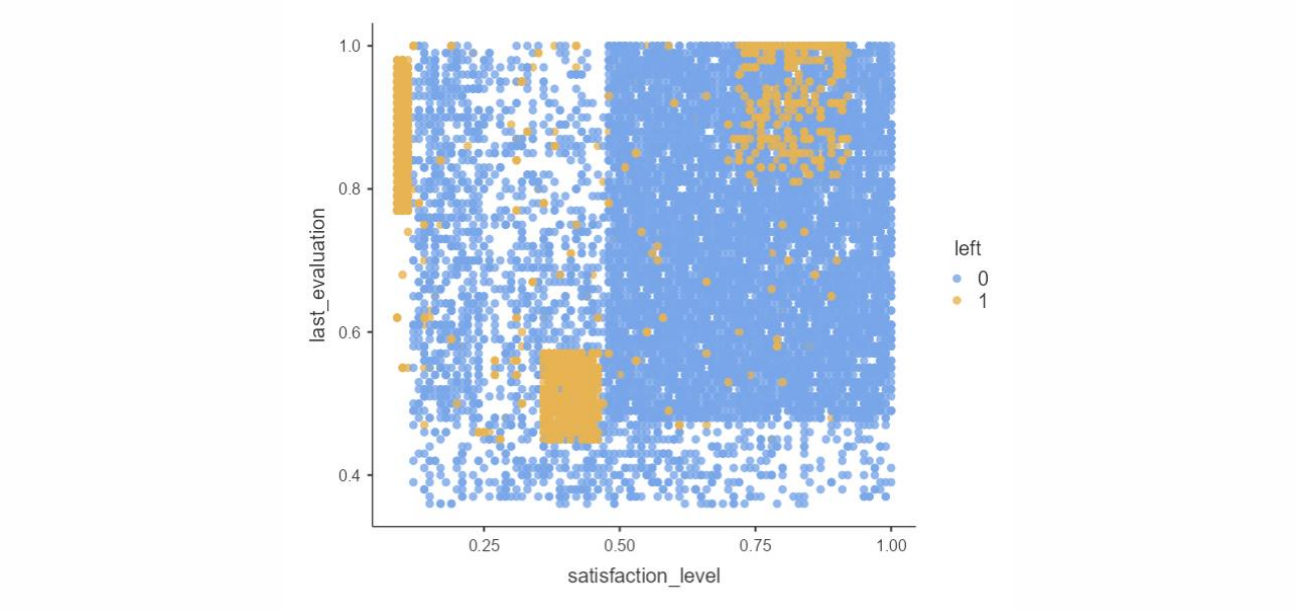



*Figure 6: Scatterplot of satisfaction level vs last evaluation grouped by left*

This scatterplot shows 3 distinct clusters for turnover i.e.

- **Sad but hard worker**: With below 0.2 satisfaction level but more than 0.75 evaluated employees left the company because though they were god at their job, they didn't feel any good at their workplace.
- **Sad and unsatisfactory worker**: This is the cluster of people who had 0.35 to 0.45 level of satisfaction and evaluated around 0.58.
- **Happy and hard worker**: These people loved their job and also had a good performance in it.

**time_spend_company:**

Descriptives

| | time_spend_company |
|---|---|
| N | 14999 |
| Missing | 0 |
| Mean | 3.50 |
| Median | 3 |
| Standard deviation | 1.46 |

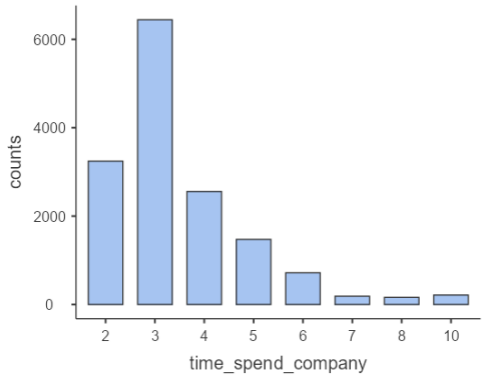



*Figure 7: Bar plot of Years at the company spent*

Most of the employees spent around 3.5 years in this company.

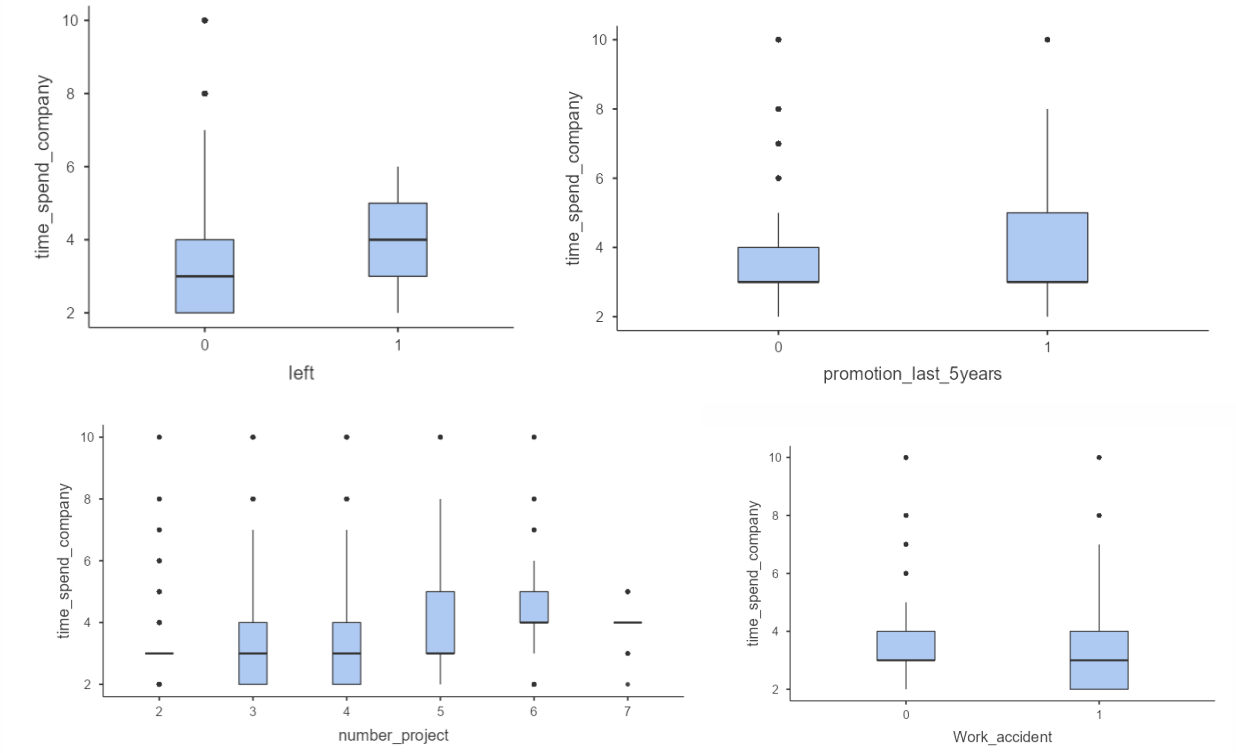



*Figure 8: Boxplot of "time_spent_company vs left", "time_spent_company vs promotion_last_5years", "time_spent_company vs number_project", and "time_spent_company vs work_accident"*

According to the above boxplots time spend at the company looks like a categorical value:

- 3 years or less
- Between 3 and 5
- More than 5

The employee under category 3 to 5 years can leave the company depending on the fact if he does not get the promotion and if he is working on 3 projects or more.

**Work_accident:**

**Frequencies**

Frequencies of Work_accident

| Levels | Counts | % of Total | Cumulative % |
|---|---|---|---|
| 0 | 12830 | 85.5 % | 85.5 % |
| 1 | 2169 | 14.5 % | 100.0 % |

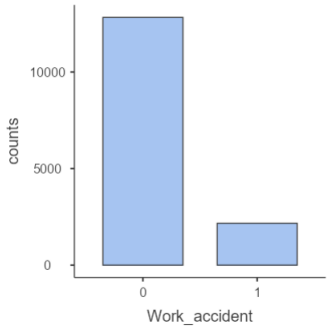



*Figure 9: Frequencies and Bar plot of Work_accident*

Frequencies of Work_accident

| | left | |
|---|---|---|
| Work_accident | 0 | 1 |
| 0 | 9428 | 3402 |
| 1 | 2000 | 169 |

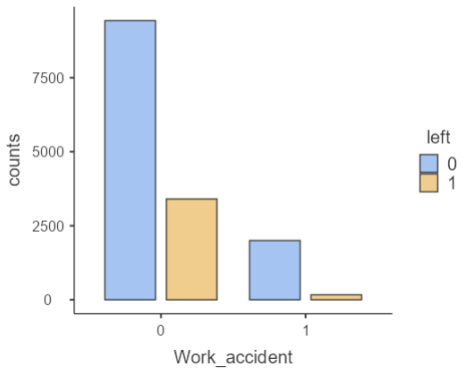



*Figure 10: Bar plot of Left vs Work_accident*

14.5% of the employees had work accident which is 2169 of total counts. Among those only 169 people left the company. It shows work accident cannot be a factory of leaving the company.

## Promotion_last_5years:

Descriptives

| | promotion_last_5years |
|---|---|
| N | 14999 |
| Missing | 0 |
| Mean | 0.0213 |

**Frequencies**

Frequencies of promotion_last_5years

| Levels | Counts | % of Total | Cumulative % |
|---|---|---|---|
| 0 | 14680 | 97.9 % | 97.9 % |
| 1 | 319 | 2.1 % | 100.0 % |

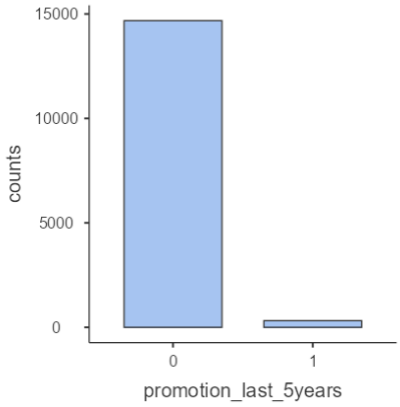



*Figure 11: Descriptive and Bar plot of promotion_last_5years*

Frequencies of promotion_last_5years

| | left | |
|---|---|---|
| promotion_last_5years | 0 | 1 |
| 0 | 11128 | 3552 |
| 1 | 300 | 19 |

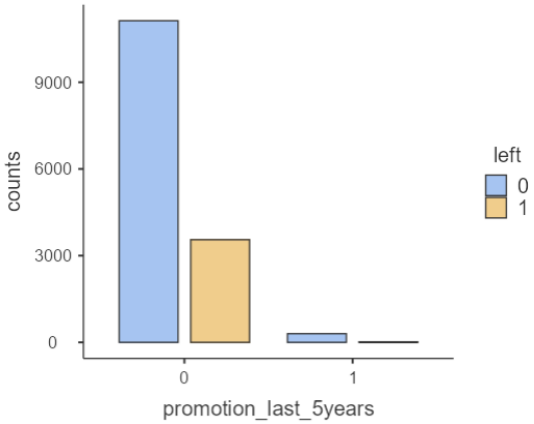



*Figure 12: Frequencies and Bar plot of promotion_last_5years vs left*

Only 2.1% of the total employees didn't get the promotion in last 5 years. A promotion should be an important factor for leaving the company. However, among those 2.1% i.e. 319 stuffs, only 19 stuffs left the company. It says that the relation between promotion and turnover is significantly lower.

## Department:

Frequencies of Department

| Levels | Counts | % of Total | Cumulative % |
|---|---|---|---|
| IT | 1227 | 8.2 % | 8.2 % |
| RandD | 787 | 5.2 % | 13.4 % |
| accounting | 767 | 5.1 % | 18.5 % |
| hr | 739 | 4.9 % | 23.5 % |
| management | 630 | 4.2 % | 27.7 % |
| marketing | 858 | 5.7 % | 33.4 % |
| product_mng | 902 | 6.0 % | 39.4 % |
| sales | 4140 | 27.6 % | 67.0 % |
| support | 2229 | 14.9 % | 81.9 % |
| technical | 2720 | 18.1 % | 100.0 % |

*Figure 13: Frequency Distribution of Department*

```
#Department

# Types of colors
color_types = ['#78C850','#F08030','#6890F0','#A8B820','#A8A878','#A040A0','#F8D030',
               '#E0C068','#EE99AC','#C03028','#F85888','#B8A038','#705898','#98D8D8','#7038F8']

# Count Plot (a.k.a. Bar Plot)
sns.countplot(x='Department', data=df, palette=color_types).set_title('Distribution of Department');

# Rotate x-labels
plt.xticks(rotation=-45)
```

```
(array([0, 1, 2, 3, 4, 5, 6, 7, 8, 9]),
 <a list of 10 Text major ticklabel objects>)
```

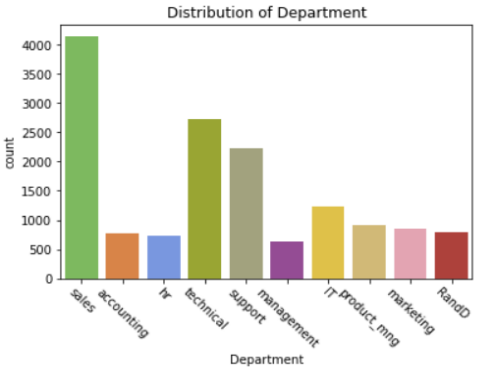
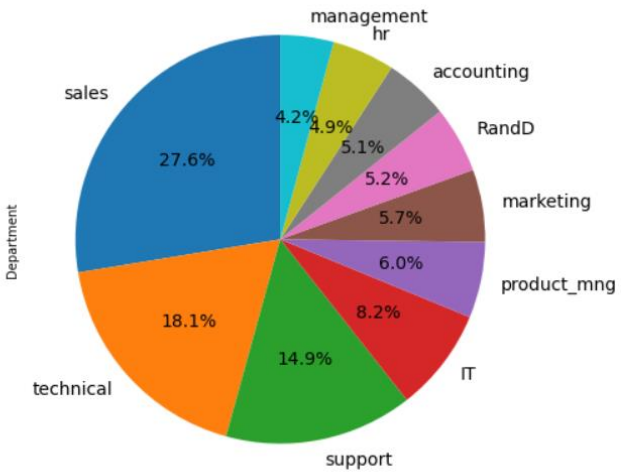



*Figure 14: Distribution of Department*

```
[19] #separate the category of Department Feature:
     Department= df['Department'].value_counts().reset_index()
     cat1 =Department['index'].to_numpy()

     # plot pie
     plt.figure(figsize=(16,8))
     ax1 = plt.subplot(121, aspect='equal')
     Department.plot(kind='pie',y = 'Department',ax=ax1,autopct='%1.1f%%', startangle=90, shadow=False, labels=cat1, legend = False, fontsize=14)

     # plot table
     from pandas.plotting import table
     ax2 = plt.subplot(122)
     plt.axis('off')
     tbl = table(ax2, Department, loc='center')
     tbl.auto_set_font_size(False)
     tbl.set_fontsize(10)
     plt.show()
```




| | index | Department |
|---|---|---|
| 0 | sales | 4140 |
| 1 | technical | 2720 |
| 2 | support | 2229 |
| 3 | IT | 1227 |
| 4 | product_mng | 902 |
| 5 | marketing | 858 |
| 6 | RandD | 787 |
| 7 | accounting | 767 |
| 8 | hr | 739 |
| 9 | management | 630 |

*Figure 15: Pie chart of Department Distribution*

```
[17] f, ax = plt.subplots(figsize=(15, 5))
     sns.countplot(y="Department", hue='left', data=df).set_title('Distribution of Department vs Left');
```

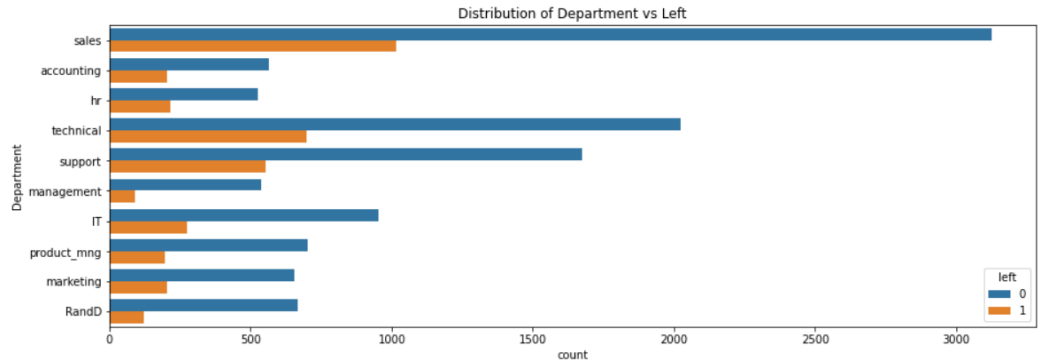



*Figure 16: Department vs left*

The top three departments are “sales”, “technical” and “support” that have employees to leave the company.

**Salary:**

Frequencies of salary

| Levels | Counts | % of Total | Cumulative % |
|---|---|---|---|
| high | 1237 | 8.2 % | 8.2 % |
| low | 7316 | 48.8 % | 57.0 % |
| medium | 6446 | 43.0 % | 100.0 % |

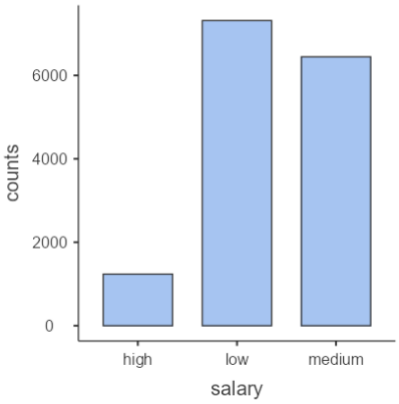



*Figure 17: Frequency Distribution of Salary*

```
[14] #salary vs left
     f, ax = plt.subplots(figsize=(15, 4))
     sns.countplot(y="salary", hue='left', data=df).set_title('Distribution of Salary vs Turnover');
```

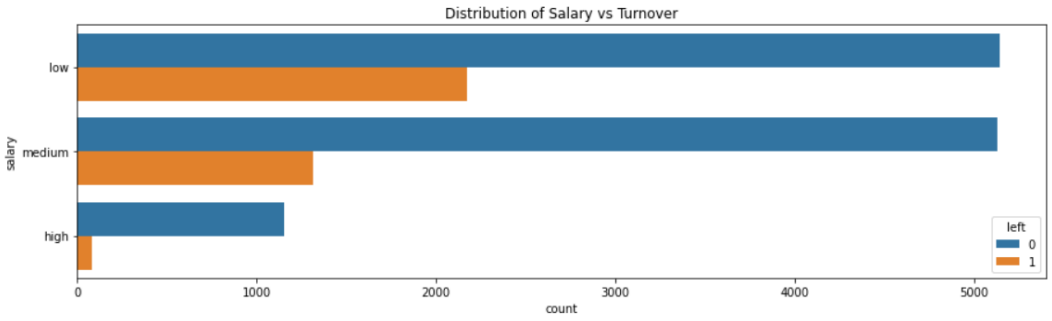



*Figure 18:Distribution of  Salary vs Left*

Salary could be a very big factor for employee turnover. It is seen that employees with low or medium salaries are more likely to leave the company.

### 1.3.2 Data Pre-Processing

Data pre-processing is important for analysing and proving the hypothesis. It includes handling missing values, detecting outliner, all the cleaning and normalizing the data.

### 1.3.3 Handling Missing Values

```
[20] #handling mising values
     df.isnull().mean()

satisfaction_level       0.0
last_evaluation          0.0
number_project           0.0
average_montly_hours     0.0
time_spend_company       0.0
Work_accident            0.0
left                     0.0
promotion_last_5years    0.0
Department               0.0
salary                   0.0
dtype: float64
```

*Figure 19: Number of missing values in the dataset*

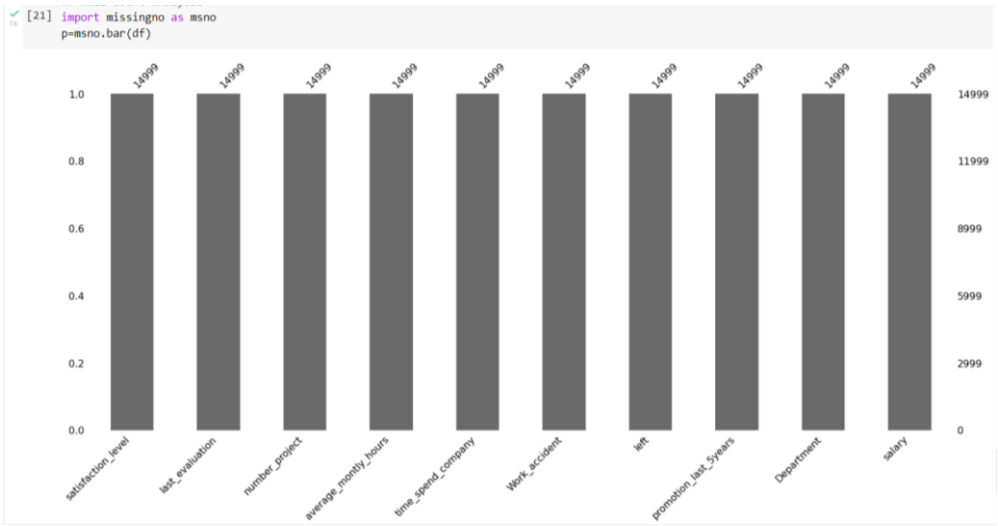



*Figure 20: Null count Analysis -Bar plot*

There is no Nan Value or missing value in the dataset. So dropping the null value process is skipped.

### 1.3.4 Handling Outliner Value

```
num= ['last_evaluation','average_montly_hours','satisfaction_level']
from scipy import stats
for col in num:
    plt.figure(figsize=(15,4))
    plt.subplot(131)
    sns.distplot(df[col], label="skew: " + str(np.round(df[col].skew(),2)))
    plt.legend()
    plt.subplot(132)
    sns.boxplot(df[col])
    plt.subplot(133)
    stats.probplot(df[col], plot=plt)
    plt.tight_layout()
    plt.show()
```

*Figure 21: Python script of handling outliner value*

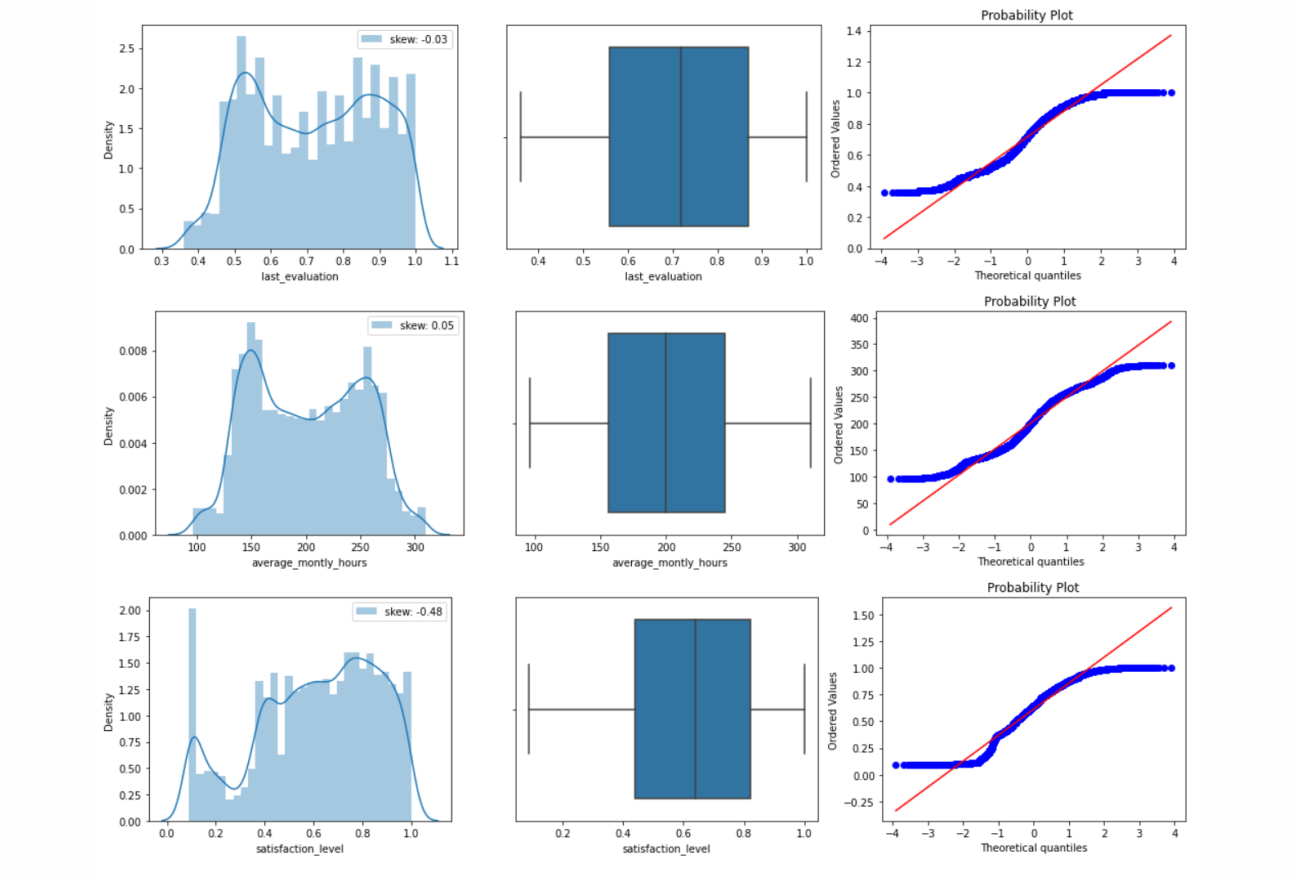



*Figure 22: Detecting outliner*

Two histograms are double-peaked. And the box plot says distinctly that there is no outliner. The q-q- plots show 45 degrees which means Gaussian distribution is maintained. To conclude, there is no outliner.

### 1.3.5 Data Cleaning

```
print('Before drop duplicate:', df.shape)

df= df.drop_duplicates()
print('After drop duplicate:', df.shape)
```

```
Before drop duplicate: (14999, 10)
After drop duplicate: (11991, 10)
```

*Figure 23: Dropping the duplicates*

There were some duplicate entries i.e. 3008 values which are dropped for cleaning purpose. So current number of employee record is 11991.

### 1.3.6 Data Normalization

There are two categorical columns i.e. salary and department. These needs to be converted as numerical. So, this dataset is normalized with one hot encoding.

```
[43] from sklearn.preprocessing import LabelEncoder
     le=LabelEncoder()
     df['salary']=le.fit_transform(df['salary'])
```

```
[46] df_new = pd.get_dummies(df,drop_first=True)
     df_new.head()
```

| | satisfaction_level | last_evaluation | number_project | average_montly_hours | time_spend_company | Work_accident | left | promotion_last_5years | sal |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.38 | 0.53 | 2 | 157 | 3 | 0 | 1 | 0 | |
| 1 | 0.80 | 0.86 | 5 | 262 | 6 | 0 | 1 | 0 | |
| 2 | 0.11 | 0.88 | 7 | 272 | 4 | 0 | 1 | 0 | |
| 3 | 0.72 | 0.87 | 5 | 223 | 5 | 0 | 1 | 0 | |
| 4 | 0.37 | 0.52 | 2 | 159 | 3 | 0 | 1 | 0 | |

```
[45] df_new.columns
```

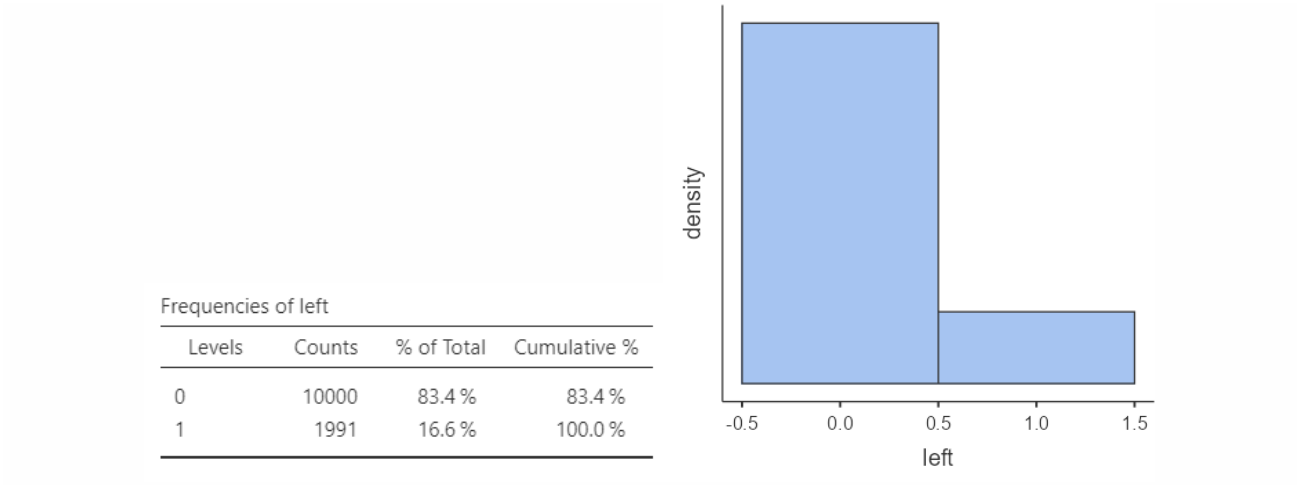
```
Index(['satisfaction_level', 'last_evaluation', 'number_project',
       'average_montly_hours', 'time_spend_company', 'Work_accident', 'left',
       'promotion_last_5years', 'salary', 'Department_RandD',
       'Department_accounting', 'Department_hr', 'Department_management',
       'Department_marketing', 'Department_product_mng', 'Department_sales',
       'Department_support', 'Department_technical'],
      dtype='object')
```

*Figure 24: One hot encoding of the dataset*

With one hot encoding, salary is replaced like high = 0, low = 1, medium = 2 and department is converted into nine different columns with 0 or 1 values. This converted data frame is saved as “featured.csv” and loaded into Jamovi software for further analysis.

```
data_path = "/content/drive/My Drive/BigDataAnalytics/"
df_new.to_csv(data_path + "featured.csv", index=False)
```

*Figure 25: Saving dataset for using in Jamovi*



Frequencies of left

| Levels | Counts | % of Total | Cumulative % |
|---|---|---|---|
| 0 | 10000 | 83.4 % | 83.4 % |
| 1 | 1991 | 16.6 % | 100.0 % |



*Figure 26: Employee turnover (16.6%) after processing the dataset*

According to this turnover rate (16.6%), still the question remains "why are the employees leaving?" And to the reach the goal of this analysis i.e. finding the factors of this question, **hypothesis testing** should be done for the mentioned research questions.

### 1.3.7 Hypothesis Testing

Hypothesis is a premise or claim that need to be tested. The terminologies that are very important regarding the hypothesis testing are – Null Hypothesis ($H_0$), Alternative Hypothesis ($H_a$ or $H_1$), Level of Significance (α), Critical Value (C), Test Statistic (t), and p-value (p).

### 1.3.8 Steps of Hypothesis Testing

To do the hypothesis testing the following figure shows the steps that need to be followed.

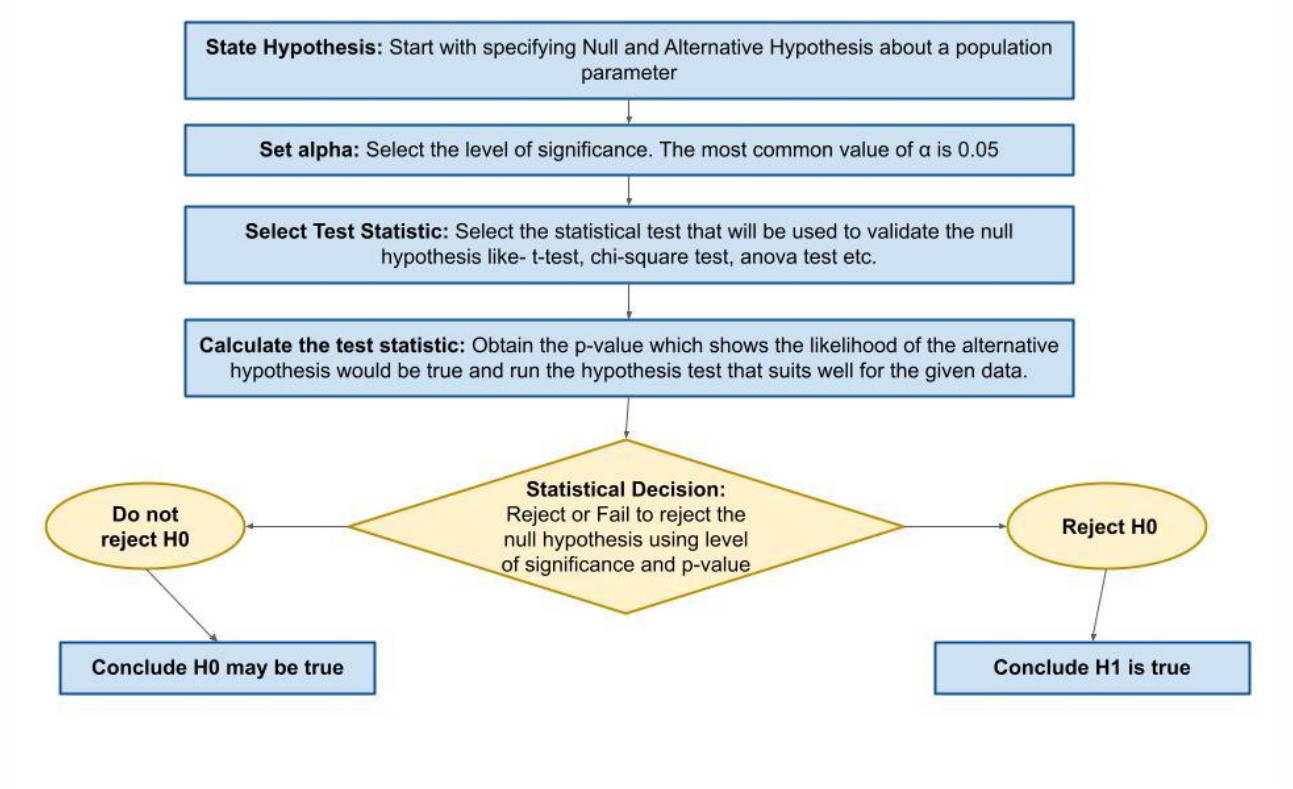



*Figure 27: Steps of Hypothesis Testing*

### 1.3.9 Hypothesis Testing of the Research Question#1

"Is "salary" the factor for making the employee leave the company?"

**Step-1:**

**$H_0$:** The factor salary is not the reason for an employee to leave the company.

**$H_1$:** An employee can leave because of the low salary.

**Step-2:** The level of significance (α) is 0.05.

**Step-3:** Since the question depends on one numerical value "salary", T-test should be performed.

**Step-4:** After doing the independent samples T-Test in Jamovi, the obtained p-value is 0.644.

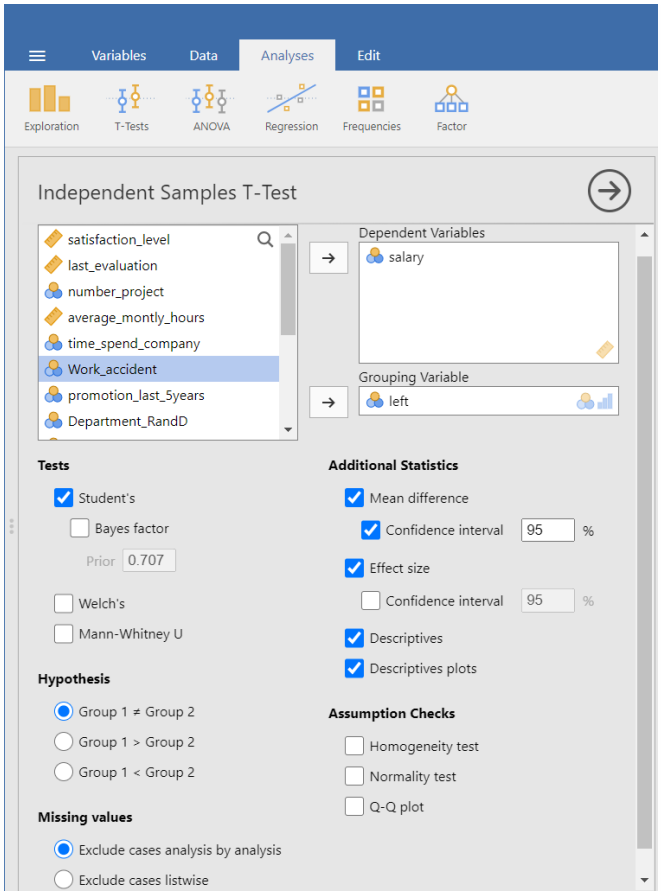



*Figure 28: Configurations of T-Test for RQ#1*

## Independent Samples T-Test

Independent Samples T-Test

| | | | | | | | 95% Confidence Interval | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Statistic | df | p | Mean difference | SE difference | Lower | Upper | | Effect Size |
| salary | Student's t | −0.463 [a] | 11989 | 0.644 | −0.00713 | 0.0154 | −0.0373 | 0.0231 | Cohen's d | −0.0114 |

[a] Levene's test is significant (p < .05), suggesting a violation of the assumption of equal variances

Group Descriptives

| | Group | N | Mean | Median | SD | SE |
|---|---|---|---|---|---|---|
| salary | 0 | 10000 | 1.35 | 1.00 | 0.646 | 0.00646 |
| | 1 | 1991 | 1.36 | 1.00 | 0.529 | 0.0118 |

*Figure 29: Statistical T-Test of RQ#1*

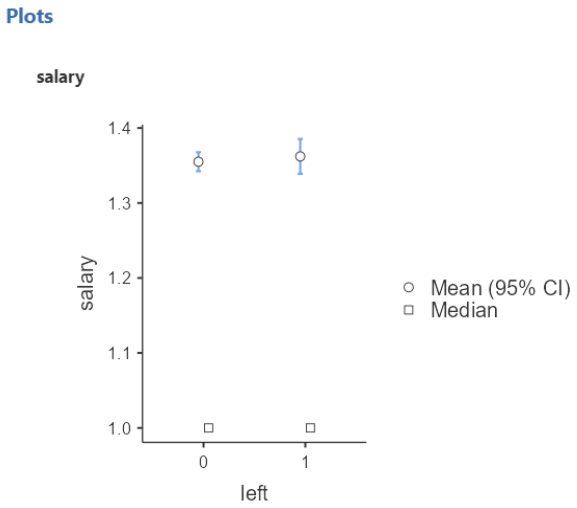



*Figure 30: Descriptive plot of T-Test for RQ#1*

There is no significant impact of salary in turnover because p-value is 0.644 which is greater than 0.05, the level of significance. The 95% confidence level shows negative in one side and positive on the other side which consistently says that it is not significant. The effect size shows Cohen's d which is the standard deviation, -0.0114. It is very close to 0. The group descriptive shows the mean median, SD and SE. The plot shows the confidence interval that says vertical line for one group is overlapping with the mean of the group. So, there is no significantly impact of "salary".

**Step-5:** As p-value is not less than alpha, the null hypothesis cannot be rejected. In conclusion, the null hypothesis i.e. "salary is not the reason to leave the company" may be true.

### 1.3.10 Hypothesis Testing of the Research Question #2

Is "satisfaction" the most important atrribute for making the employee leave the company?

**Step-1:**

**$H_0$:** The factor "satisfaction"is not the reason for an employee to leave the company.

**$H_1$:** An employee can leave because he/she is not satisfied with his/her job.

**Step-2:** The level of significance (α) is 0.05.

**Step-3:** Since the question depends on one numerical value "satisfaction", T-test should be performed.

**Step-4:** After doing the independent samples T-Test in Jamovi, the obtained p-value is <0.001.

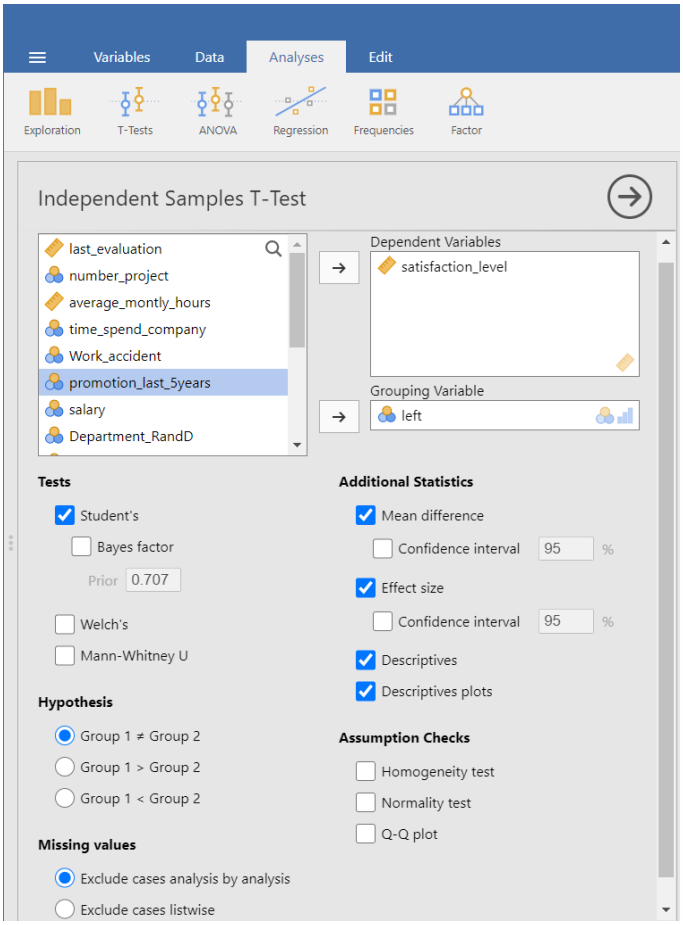



*Figure 31: Configurations of T-Test for RQ#2*

## Independent Samples T-Test

Independent Samples T-Test

| | | | | | | | 95% Confidence Interval | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Statistic | df | p | Mean difference | SE difference | Lower | Upper | | Effect Size |
| satisfaction_level | Student's t | 41.0 [a] | 11989 | < .001 | 0.227 | 0.00554 | 0.216 | 0.238 | Cohen's d | 1.01 |

[a] Levene's test is significant (p < .05), suggesting a violation of the assumption of equal variances

Group Descriptives

| | Group | N | Mean | Median | SD | SE |
|---|---|---|---|---|---|---|
| satisfaction_level | 0 | 10000 | 0.667 | 0.690 | 0.217 | 0.00217 |
| | 1 | 1991 | 0.440 | 0.410 | 0.265 | 0.00594 |

*Figure 32: Statistical T-Test of RQ#2*

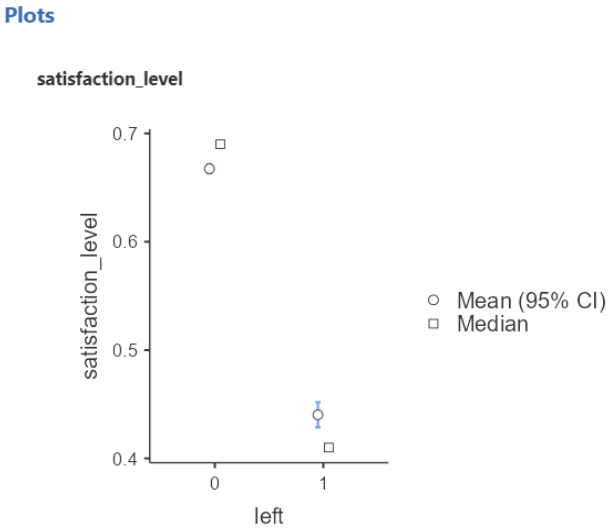



*Figure 33: Descriptive plots of T-Test for RQ#2*

There is significant impact of satisfaction level in turnover because p-value is <0.001 which is less than 0.05, the level of significance. The 95% confidence level shows similar value on both sides which consistently says that it is significant. The effect size shows Cohen's d which is the standard deviation1.01. It is definitely not close to 0. The group descriptive shows the mean median, SD and SE. The plot shows the confidence interval that says vertical line for one group is far away from the mean of the group. So, there is significantly impact of "salary".

**Step-5:** As p-value is less than alpha, the null hypothesis can be rejected. In conclusion, the alternative hypothesis i.e. "An employee can leave because he/she is not satisfied with his/her job" is true.

### 1.3.11 Hypothesis Testing of the Research Question #3

Do "the number of project" and "evaluation" impact together on turnover of the company?

**Step 1:**

**$H_0$:** Number of project and evalution do not impact together on turnover of the company.

**$H_1$:** Number of project and evalution impact together on turnover of the company.

**Step 2:** The level of significance (α) is 0.05.

**Step 3:** Since the question depends on two numerical value "number_project" and "last_evalution", first correlation should be checked and then T-test should be performed.

**Step 4:**

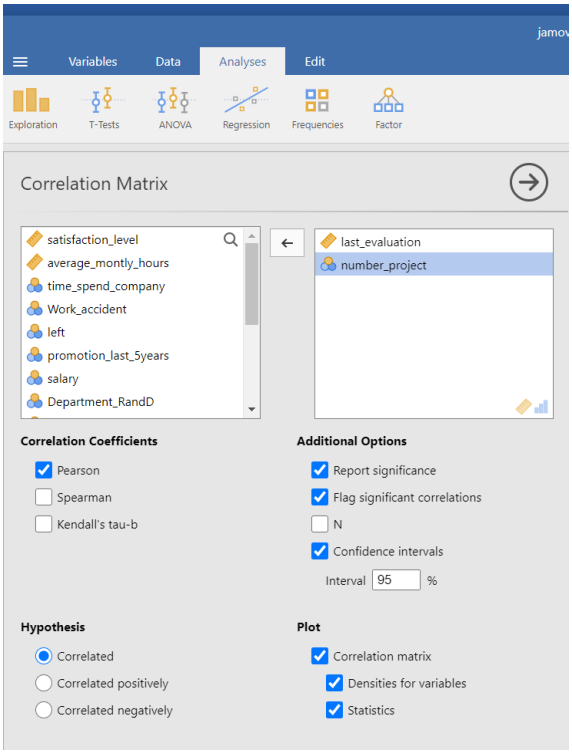
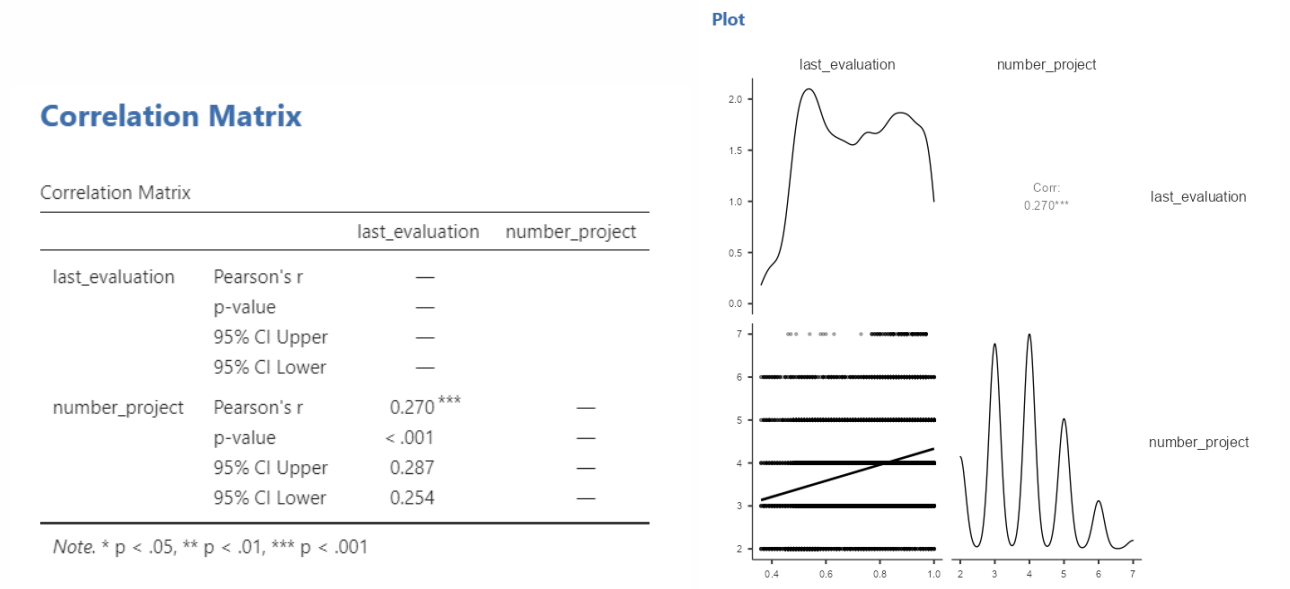



*Figure 34:Configuration of Correlation Matrix and plots for number of project and last evaluation*



**Correlation Matrix**

Correlation Matrix

| | | last_evaluation | number_project |
|---|---|---|---|
| last_evaluation | Pearson's r | — | |
| | p-value | — | |
| | 95% CI Upper | — | |
| | 95% CI Lower | — | |
| number_project | Pearson's r | 0.270 *** | — |
| | p-value | < .001 | — |
| | 95% CI Upper | 0.287 | — |
| | 95% CI Lower | 0.254 | — |

*Note.* * p < .05, ** p < .01, *** p < .001



*Figure 35: Correlation Matrix of last_evaluation and number_project*

Last evaluation and number of project handled by the employee is highly correlated because the value is 0.270 and the p-value I less than 0.001.

**Paired Samples T-Test**

Paired Samples T-Test

| | | | | | | | | 95% Confidence Interval | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | statistic | df | p | Mean difference | SE difference | Lower | Upper | | Effect Size |
| last_evaluation | number_project | Student's t | −299 | 11990 | < .001 | −3.09 | 0.0103 | −3.11 | −3.07 | Cohen's d | −2.73 |

Descriptives

| | N | Mean | Median | SD | SE |
|---|---|---|---|---|---|
| last_evaluation | 11991 | 0.717 | 0.720 | 0.168 | 0.00154 |
| number_project | 11991 | 3.803 | 4 | 1.163 | 0.01062 |

*Figure 36: Paired Sampled T-test for RQ#3*

**Plots**

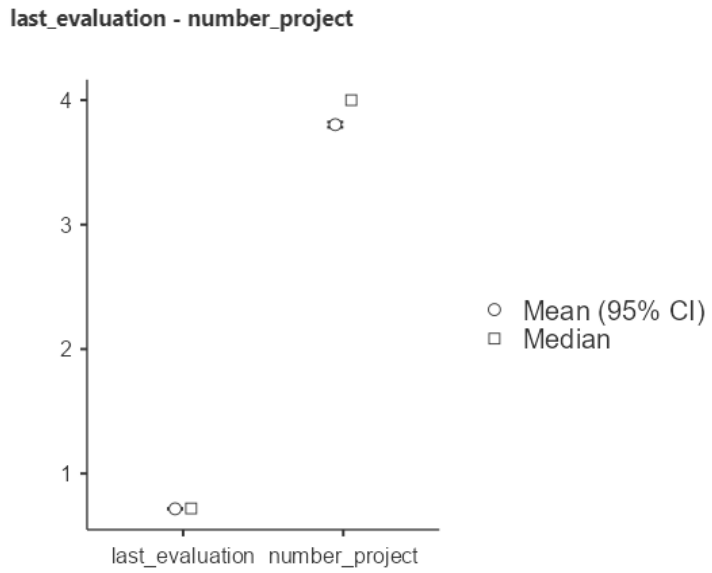



*Figure 37: Distribution plot of t-test for RQ#3*

The paired sampled t-test of last_evaluation and number_project shows p-values is less than 0.01. And the distribution plot did not overlap. So, these two variables singnificantly impacts on turnover together.

**Step-5:** p-value is less than the alpha. So. Null hypothesis is rejected. So alternative hypothesis "Number of project and evalution impact together on turnover of the company" is true.

## 1.4 Task-4: Evaluation and Future Development

### 1.4.1 Evaluation and Recommendations

1. The most important attribute is “satisfaction” for employee turnover. The employers of this company should give attention to the low satisfied employees with around 0.440 and try to increase their satisfaction level.
2. According to the t-test salary does not effect on the turnover. However, exploratory analysis says that employees with low or medium salaries are more likely to leave the company. So, it is better to keep in mind to satisfy the employees with at least average salary so that they don’t leave the company.
3. The number of project” and “evaluation” impact together on turnover of the company. So, from analysis, the employees who are handling around 2 projects and getting 0.5 centred evaluation need attention from the company because they can cause employee turnover.
4. From the analysis, the cluster “Sad but hard worker” can leave the company. So, with below 0.2 satisfaction level but more than 0.75 evaluated employees should be paid attention by the employers.
5. According to the result, the three most important factors that impacts on turnover are – satisfaction_level, time_spend_cpany and last_evaluation. So, even with high satisfaction and high evaluation employees who spend more hours per month in the company need attention.

### 1.4.2 Limitations of the Work

1. It is an imbalanced dataset with 76:23 ratio of turnover. It is a huge limitation to conclude any decision. New technologies can be used to gather balanced data or oversampling or under sampling can be done to balance this dataset.
2. To analyse the dataset, statistical approach has been used. Other analytical approaches like machine learning or data mining could be used. Because the HR department require the skillset to work with large dataset.
3. Besides Google Cloud, other cloud services such as AWS or Microsoft Azure could be a good machine learning tool for analytical part.
4. Besides Jamovi, there are some good big data platforms like – Apache haddop, Cloudera, spark etc. Among them, Spark could be used for analysis because this dataset is based on a prediction problem and Spark is preferable when speed is prioritized over price and it deals with smaller data batches when quick analytics results are required.
5. Using modern technologies to observe and collect more data, as well as making predictions based on data, might raise ethical concerns.

### 1.4.3 Future Work

Employee turnover has a big impact on a company as well as employee’s livelihood. According to the EDA and the hypothesis testing, there are many reasons that make employee give resignation. HR department plays a huge role here. They can use these factors to **create a classifier model to predict which employee is going to leave next**.

Using a classifier to predict the turnover can prevent letting go of high performing valuable employees of the company. Instead of letting high evaluated people leave, the company can focus on their satisfaction and let worse performing employee leave. Therefore, HR department can easily save the damage to the organization compared to saving cost for a cheaper one.

To check the accuracy, **Random Forest** or **Gradient Boosting** can be some preferable algorithms which will help to create model for prediction. With the machine learning technique, a **Decision Tree** can be used to explain and give the recommendations to the company. Because with the reduced max depth and F1 score it significantly shows the dependency of the factors on turnover from which many decisions can be taken. So, using this **Decision Tree Classifier,** the most important feature can be ranked during prediction. And using **Logistics Regression** it can be visualized that which employees are most likely to leave the company.

With the odds data in hand, the responsible sector (usually HR) could draw up plans to approach these employees and ensure their performance, for instance: **career plans, financial incentives, motivational speeches, other benefits** etc.

However, The HR department should remember that there may be cases of employees with a high probability of leaving the company, but who end up not leaving after a while (**false positives**), something that can generate an unnecessary cost for the company, in addition to employees with a low probability of leaving. leave the company, but who end up leaving after a while (**false negatives**), which, as previously mentioned, can lead to the loss of talent and higher costs for new hires.

Therefore, despite of having false positive and false negative cases, the company can take advantages of new technologies like machine learning and expand the solution for further insights.

# Section 2 Business Intelligence (Tableau)

As a Data Analyst by one of the leading supermarkets in the US and provided with the dataset named “Appendix_all_sales_data.csv”. The dataset in Tableau looks like the following:

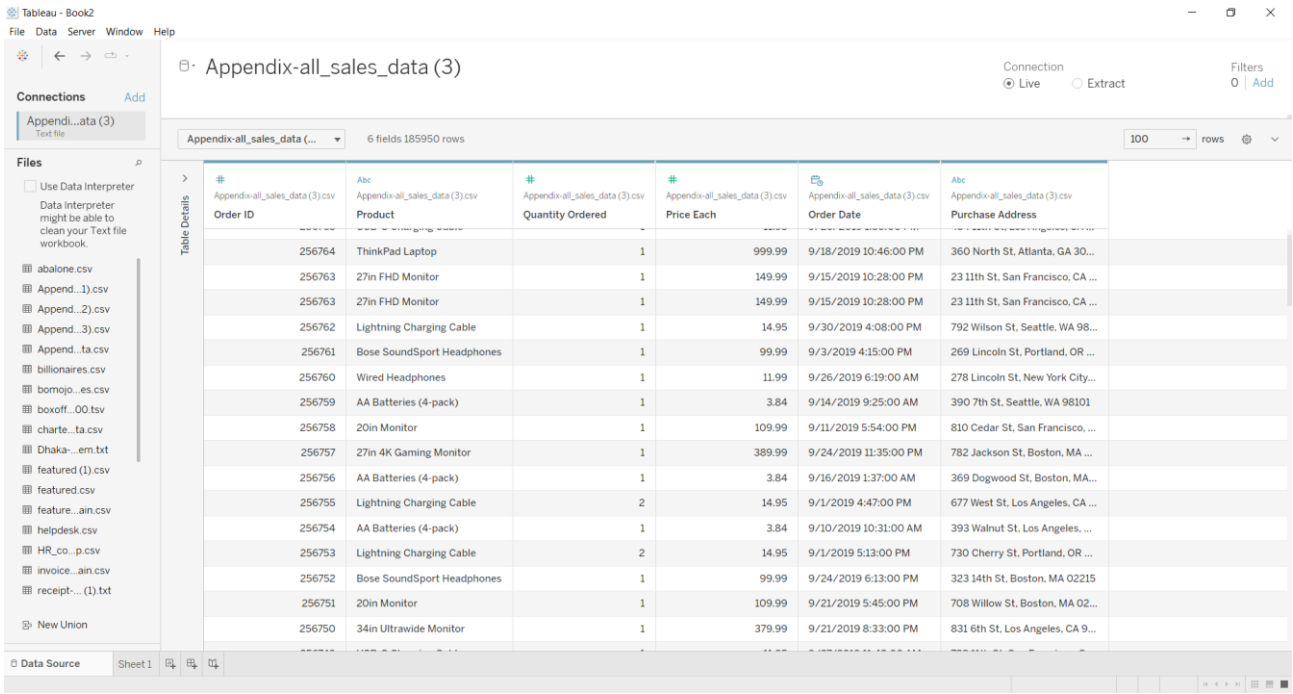
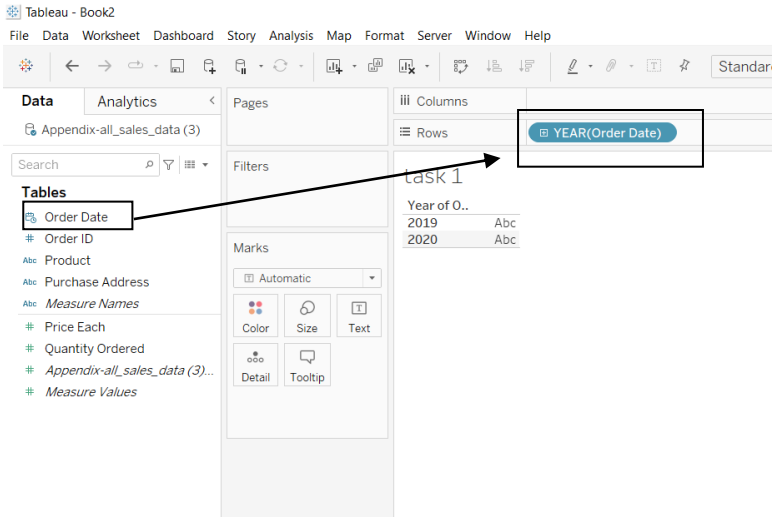



*Figure 38: The given dataset in Tableau*

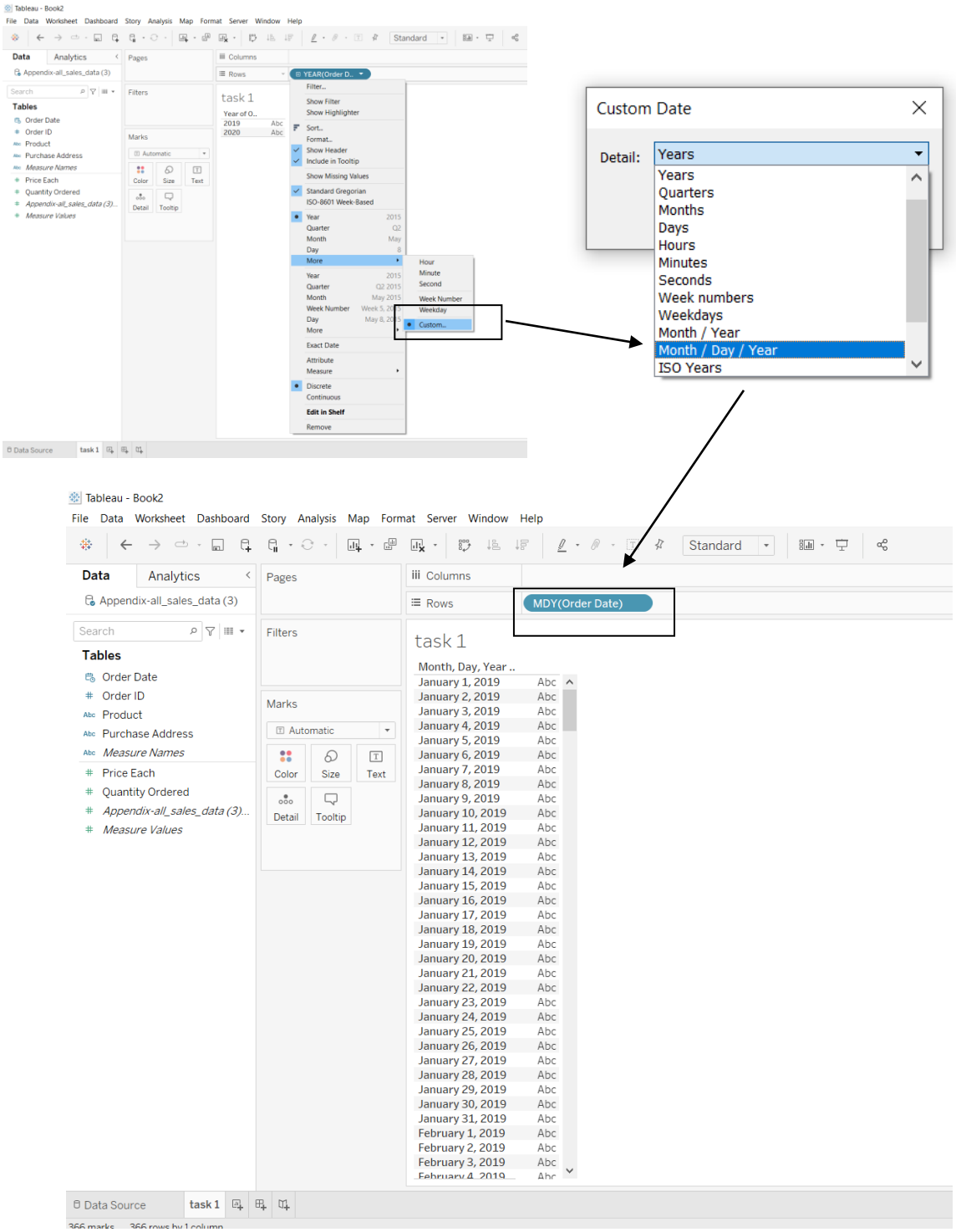
## 2.1 Task 1

Given the item ordered comes with 3 months’ warranty from the date it was ordered. Calculating the Warranty End date using a date function:

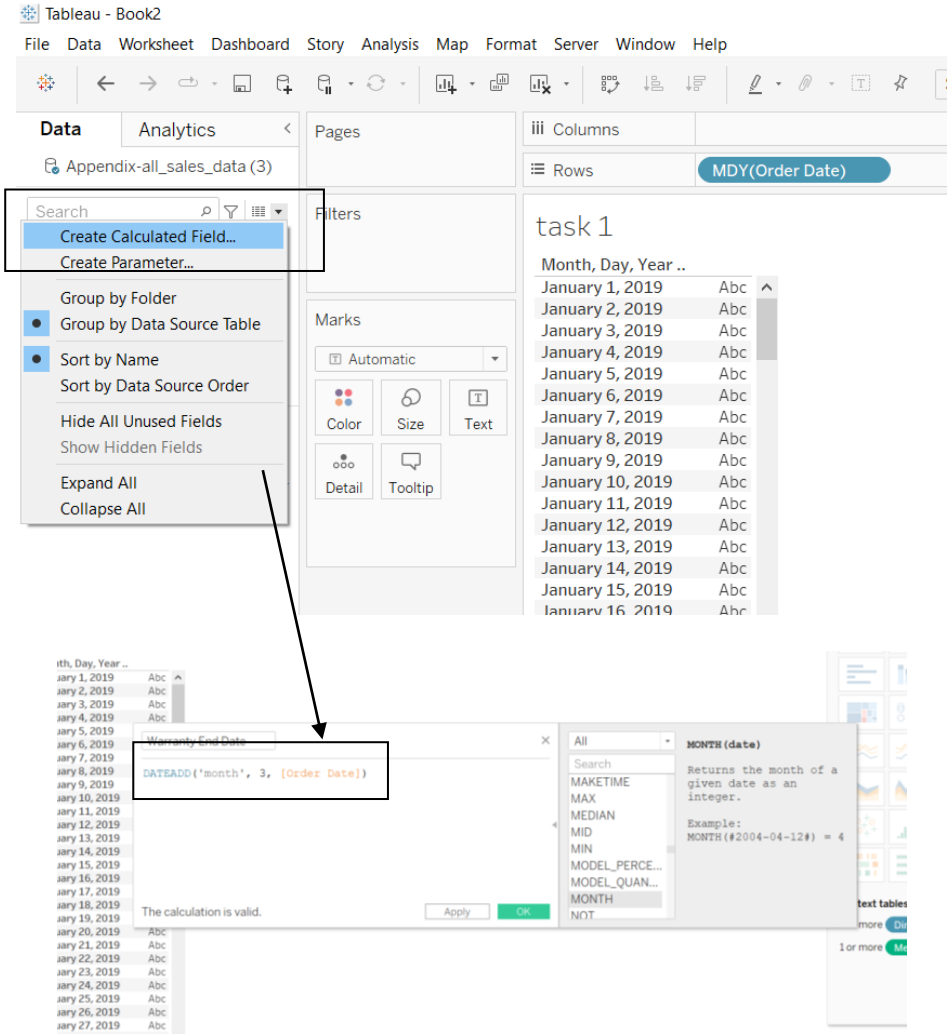
1. Order Date column is dragged to the rows section.



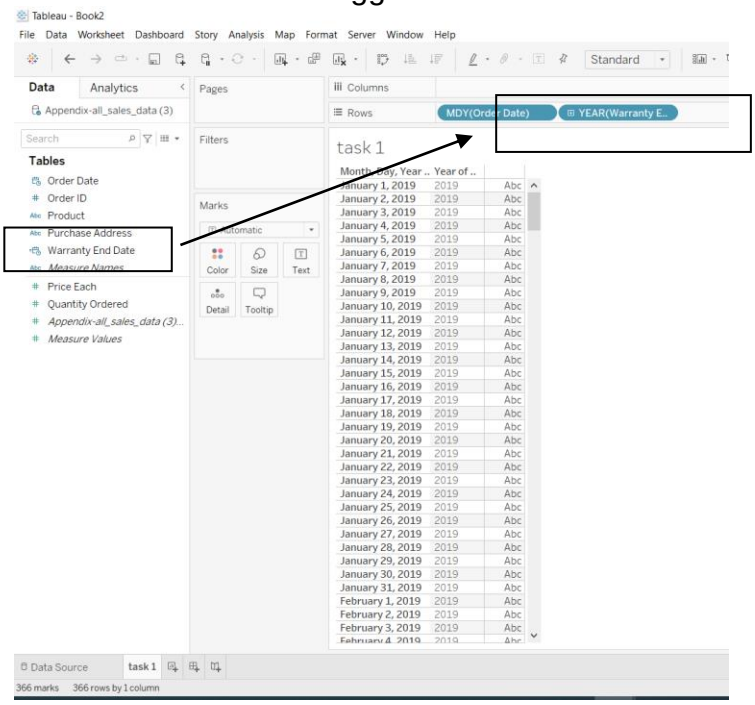

2. By right clicking on the rows, the format of the order date is changed to MDY format by choosing the custom option.



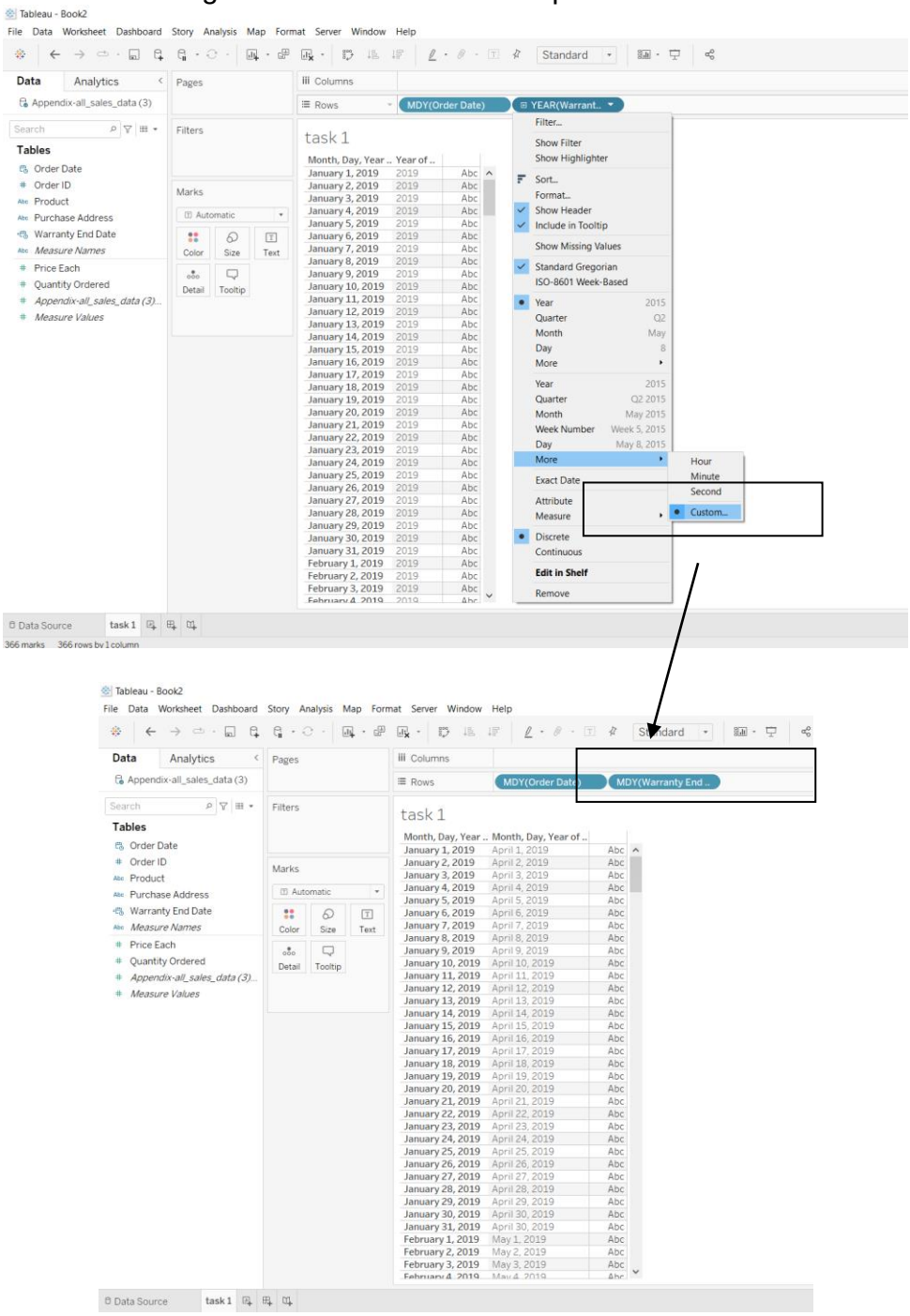

3. By choosing create calculated field, a field is created named “Warrenty End Date” using DATEADD() function that adds 3 months with each date.




4. The new field warrant end date is dragged to the rows section.



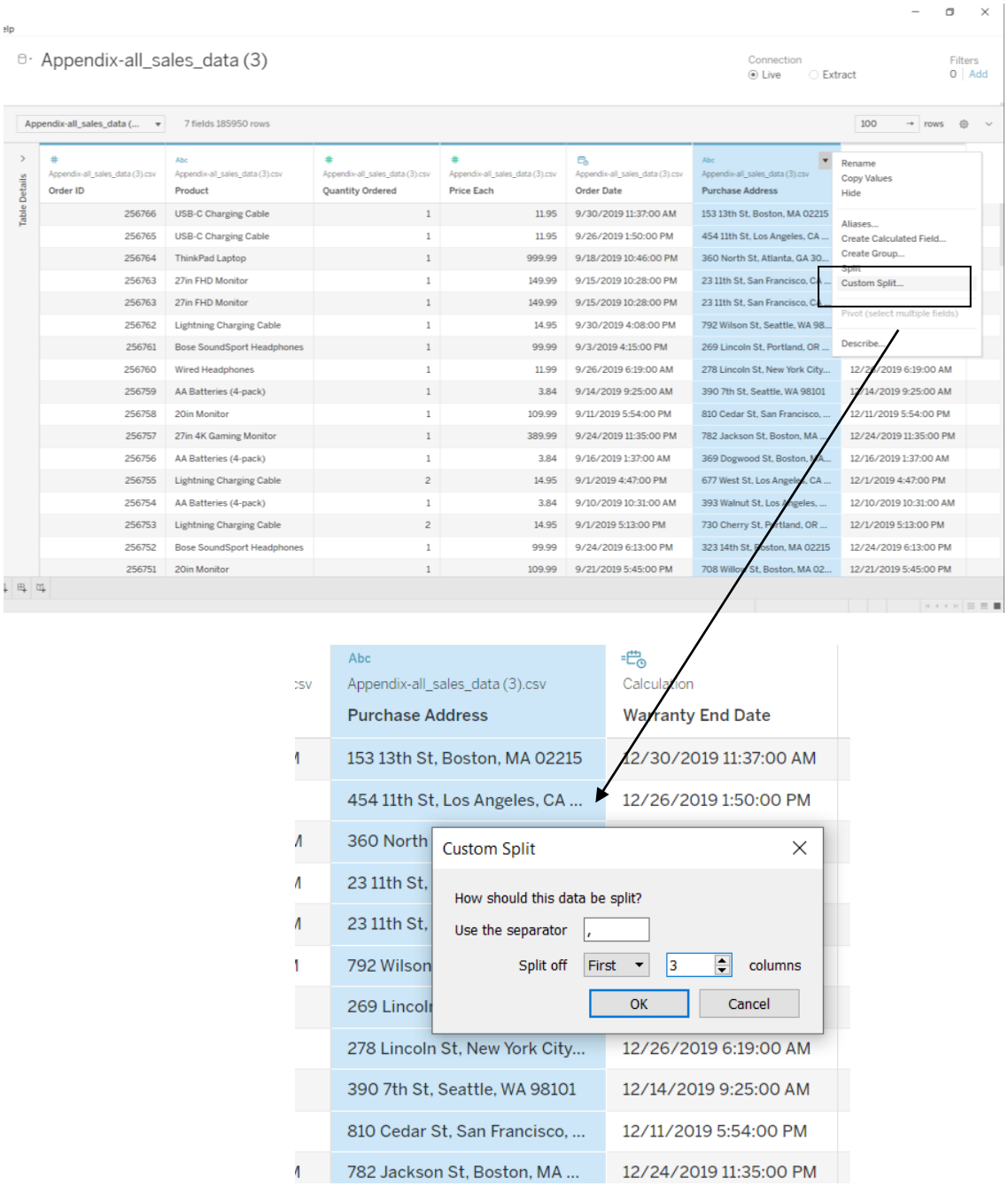
5. Date format is changed to MDY with custom option.



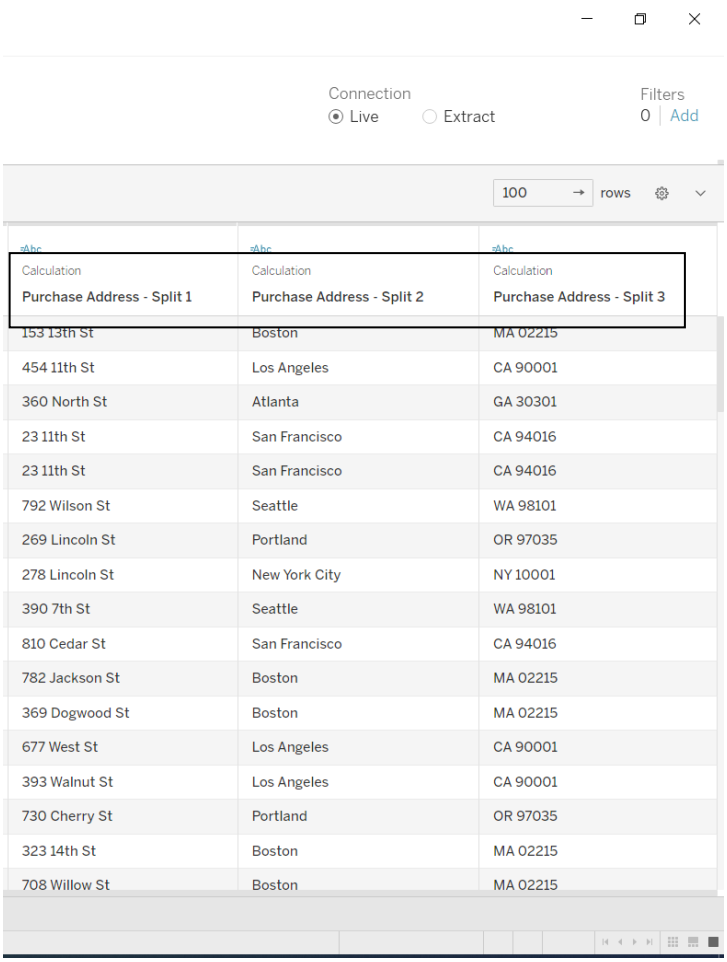

## 2.2 Task 2

Using an appropriate chart, displaying the total sales made in each City and ordering the result by the city with the highest sales by giving the user an option to either choose Sum of Sales or Average Sales.

1. Splitting the purchase address with comma separator into 3 columns using custom split.


Appendix-all_sales_data (3)
Connection
Live
Extract
Filters
0 | Add
7 fields 185950 rows
100 rows
Table Details
Order ID
Product
Quantity Ordered
Price Each
Order Date
Purchase Address
Rename
Copy Values
Hide
Aliases...
Create Calculated Field...
Create Group...
Split
Custom Split...
Pivot (select multiple fields)
Describe...
256766 USB-C Charging Cable 1 11.95 9/30/2019 11:37:00 AM 153 13th St, Boston, MA 02215
256765 USB-C Charging Cable 1 11.95 9/26/2019 1:50:00 PM 454 11th St, Los Angeles, CA ...
256764 ThinkPad Laptop 1 999.99 9/18/2019 10:46:00 PM 360 North St, Atlanta, GA 30...
256763 27in FHD Monitor 1 149.99 9/15/2019 10:28:00 PM 23 11th St, San Francisco, CA ...
256763 27in FHD Monitor 1 149.99 9/15/2019 10:28:00 PM 23 11th St, San Francisco, CA ...
256762 Lightning Charging Cable 1 14.95 9/30/2019 4:08:00 PM 792 Wilson St, Seattle, WA 98...
256761 Bose SoundSport Headphones 1 99.99 9/3/2019 4:15:00 PM 269 Lincoln St, Portland, OR ...
256760 Wired Headphones 1 11.99 9/26/2019 6:19:00 AM 278 Lincoln St, New York City... 12/26/2019 6:19:00 AM
256759 AA Batteries (4-pack) 1 3.84 9/14/2019 9:25:00 AM 390 7th St, Seattle, WA 98101 12/14/2019 9:25:00 AM
256758 20in Monitor 1 109.99 9/11/2019 5:54:00 PM 810 Cedar St, San Francisco, ... 12/11/2019 5:54:00 PM
256757 27in 4K Gaming Monitor 1 389.99 9/24/2019 11:35:00 PM 782 Jackson St, Boston, MA ... 12/24/2019 11:35:00 PM
256756 AA Batteries (4-pack) 1 3.84 9/16/2019 1:37:00 AM 369 Dogwood St, Boston, MA... 12/16/2019 1:37:00 AM
256755 Lightning Charging Cable 2 14.95 9/1/2019 4:47:00 PM 677 West St, Los Angeles, CA ... 12/1/2019 4:47:00 PM
256754 AA Batteries (4-pack) 1 3.84 9/10/2019 10:31:00 AM 393 Walnut St, Los Angeles, ... 12/10/2019 10:31:00 AM
256753 Lightning Charging Cable 2 14.95 9/1/2019 5:13:00 PM 730 Cherry St, Portland, OR ... 12/1/2019 5:13:00 PM
256752 Bose SoundSport Headphones 1 99.99 9/24/2019 6:13:00 PM 323 14th St, Boston, MA 02215 12/24/2019 6:13:00 PM
256751 20in Monitor 1 109.99 9/21/2019 5:45:00 PM 708 Willow St, Boston, MA 02... 12/21/2019 5:45:00 PM


Abc
Appendix-all_sales_data (3).csv
Purchase Address
Calculation
Warranty End Date
153 13th St, Boston, MA 02215 12/30/2019 11:37:00 AM
454 11th St, Los Angeles, CA ... 12/26/2019 1:50:00 PM
360 North
23 11th St,
23 11th St,
792 Wilson
269 Lincoln
278 Lincoln St, New York City... 12/26/2019 6:19:00 AM
390 7th St, Seattle, WA 98101 12/14/2019 9:25:00 AM
810 Cedar St, San Francisco, ... 12/11/2019 5:54:00 PM
782 Jackson St, Boston, MA ... 12/24/2019 11:35:00 PM
Custom Split
How should this data be split?
Use the separator ,
Split off First 3 columns
OK
Cancel



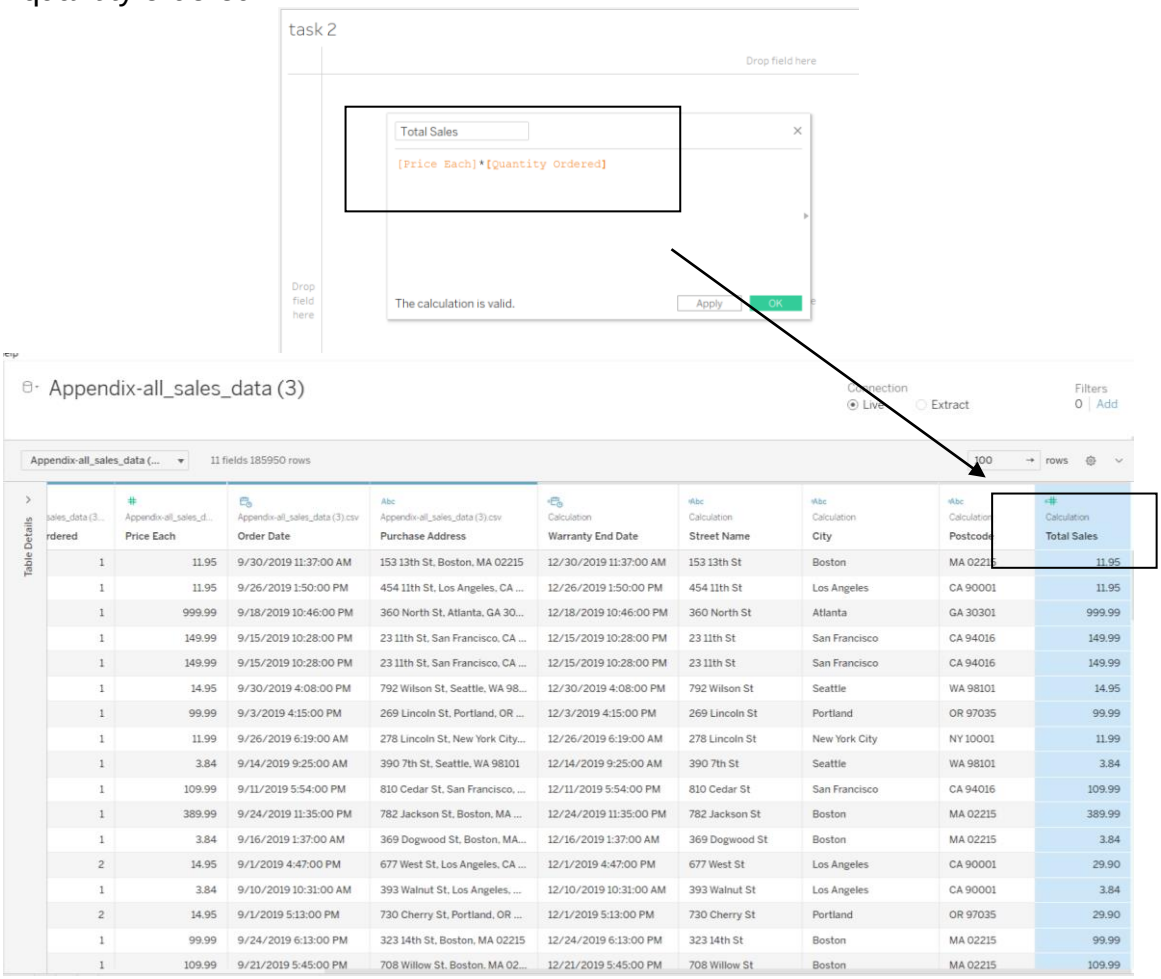

| Calculation Purchase Address - Split 1 | Calculation Purchase Address - Split 2 | Calculation Purchase Address - Split 3 |
|---|---|---|
| 153 13th St | Boston | MA 02215 |
| 454 11th St | Los Angeles | CA 90001 |
| 360 North St | Atlanta | GA 30301 |
| 23 11th St | San Francisco | CA 94016 |
| 23 11th St | San Francisco | CA 94016 |
| 792 Wilson St | Seattle | WA 98101 |
| 269 Lincoln St | Portland | OR 97035 |
| 278 Lincoln St | New York City | NY 10001 |
| 390 7th St | Seattle | WA 98101 |
| 810 Cedar St | San Francisco | CA 94016 |
| 782 Jackson St | Boston | MA 02215 |
| 369 Dogwood St | Boston | MA 02215 |
| 677 West St | Los Angeles | CA 90001 |
| 393 Walnut St | Los Angeles | CA 90001 |
| 730 Cherry St | Portland | OR 97035 |
| 323 14th St | Boston | MA 02215 |
| 708 Willow St | Boston | MA 02215 |

2. Renaming the column names to Street Name, City and Postcode.

| Calculation Street Name | Calculation City | Calculation Postcode |
|---|---|---|
| 153 13th St | Boston | MA 02215 |
| 454 11th St | Los Angeles | CA 90001 |
| 360 North St | Atlanta | GA 30301 |
| 23 11th St | San Francisco | CA 94016 |
| 23 11th St | San Francisco | CA 94016 |
| 792 Wilson St | Seattle | WA 98101 |
| 269 Lincoln St | Portland | OR 97035 |
| 278 Lincoln St | New York City | NY 10001 |
| 390 7th St | Seattle | WA 98101 |
| 810 Cedar St | San Francisco | CA 94016 |
| 782 Jackson St | Boston | MA 02215 |
| 369 Dogwood St | Boston | MA 02215 |
| 677 West St | Los Angeles | CA 90001 |
| 393 Walnut St | Los Angeles | CA 90001 |
| 730 Cherry St | Portland | OR 97035 |
| 323 14th St | Boston | MA 02215 |
| 708 Willow St | Boston | MA 02215 |

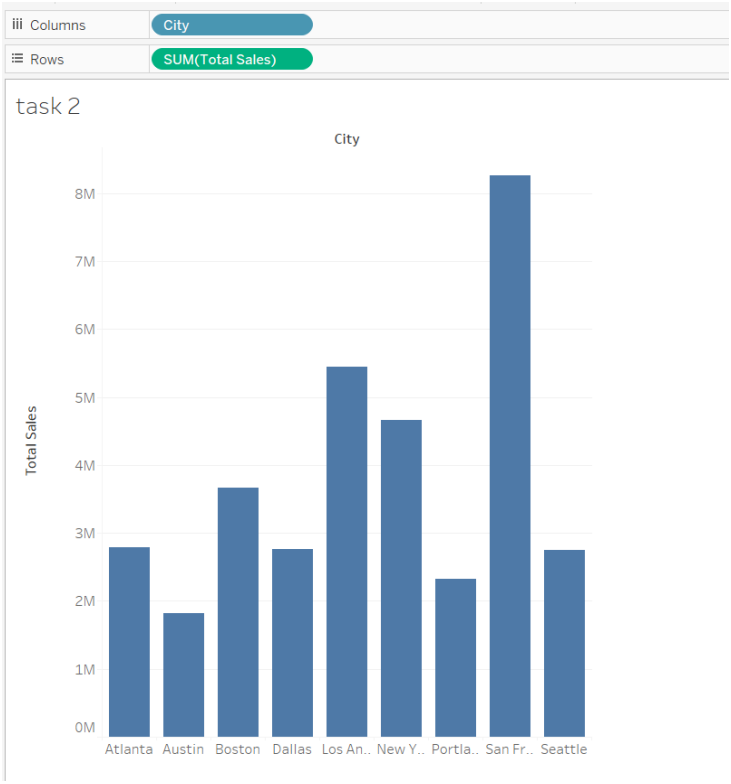
3. Creating a calculated field named “Total sales” by multiplying Price each and

quantity ordered.

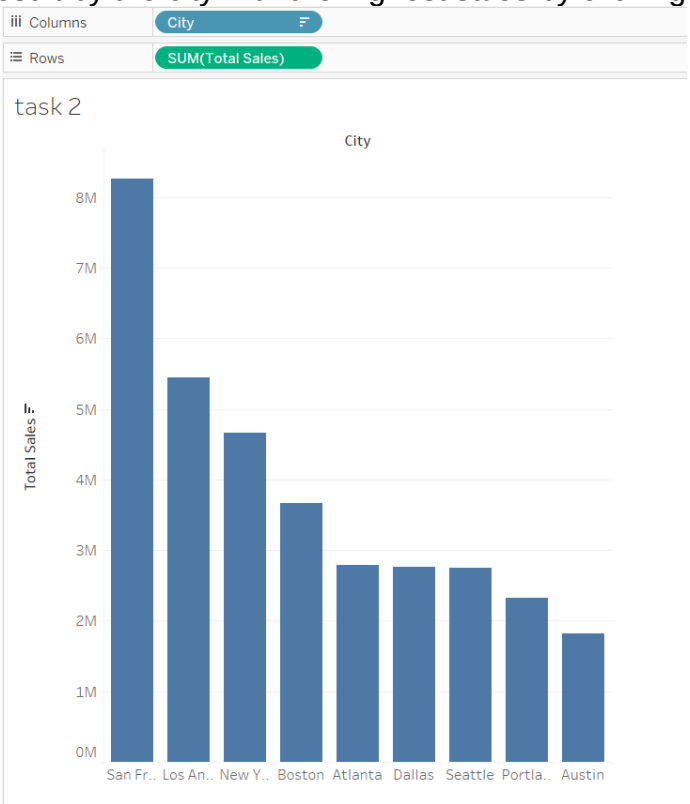



| ...rdered | Price Each | Order Date | Purchase Address | Warranty End Date | Street Name | City | Postcode | Total Sales |
|---|---|---|---|---|---|---|---|---|
| 1 | 11.95 | 9/30/2019 11:37:00 AM | 153 13th St, Boston, MA 02215 | 12/30/2019 11:37:00 AM | 153 13th St | Boston | MA 02215 | 11.95 |
| 1 | 11.95 | 9/26/2019 1:50:00 PM | 454 11th St, Los Angeles, CA ... | 12/26/2019 1:50:00 PM | 454 11th St | Los Angeles | CA 90001 | 11.95 |
| 1 | 999.99 | 9/18/2019 10:46:00 PM | 360 North St, Atlanta, GA 30... | 12/18/2019 10:46:00 PM | 360 North St | Atlanta | GA 30301 | 999.99 |
| 1 | 149.99 | 9/15/2019 10:28:00 PM | 23 11th St, San Francisco, CA ... | 12/15/2019 10:28:00 PM | 23 11th St | San Francisco | CA 94016 | 149.99 |
| 1 | 149.99 | 9/15/2019 10:28:00 PM | 23 11th St, San Francisco, CA ... | 12/15/2019 10:28:00 PM | 23 11th St | San Francisco | CA 94016 | 149.99 |
| 1 | 14.95 | 9/30/2019 4:08:00 PM | 792 Wilson St, Seattle, WA 98... | 12/30/2019 4:08:00 PM | 792 Wilson St | Seattle | WA 98101 | 14.95 |
| 1 | 99.99 | 9/3/2019 4:15:00 PM | 269 Lincoln St, Portland, OR ... | 12/3/2019 4:15:00 PM | 269 Lincoln St | Portland | OR 97035 | 99.99 |
| 1 | 11.99 | 9/26/2019 6:19:00 AM | 278 Lincoln St, New York City... | 12/26/2019 6:19:00 AM | 278 Lincoln St | New York City | NY 10001 | 11.99 |
| 1 | 3.84 | 9/14/2019 9:25:00 AM | 390 7th St, Seattle, WA 98101 | 12/14/2019 9:25:00 AM | 390 7th St | Seattle | WA 98101 | 3.84 |
| 1 | 109.99 | 9/11/2019 5:54:00 PM | 810 Cedar St, San Francisco, ... | 12/11/2019 5:54:00 PM | 810 Cedar St | San Francisco | CA 94016 | 109.99 |
| 1 | 389.99 | 9/24/2019 11:35:00 PM | 782 Jackson St, Boston, MA ... | 12/24/2019 11:35:00 PM | 782 Jackson St | Boston | MA 02215 | 389.99 |
| 1 | 3.84 | 9/16/2019 1:37:00 AM | 369 Dogwood St, Boston, MA... | 12/16/2019 1:37:00 AM | 369 Dogwood St | Boston | MA 02215 | 3.84 |
| 2 | 14.95 | 9/1/2019 4:47:00 PM | 677 West St, Los Angeles, CA ... | 12/1/2019 4:47:00 PM | 677 West St | Los Angeles | CA 90001 | 29.90 |
| 1 | 3.84 | 9/10/2019 10:31:00 AM | 393 Walnut St, Los Angeles, ... | 12/10/2019 10:31:00 AM | 393 Walnut St | Los Angeles | CA 90001 | 3.84 |
| 2 | 14.95 | 9/1/2019 5:13:00 PM | 730 Cherry St, Portland, OR ... | 12/1/2019 5:13:00 PM | 730 Cherry St | Portland | OR 97035 | 29.90 |
| 1 | 99.99 | 9/24/2019 6:13:00 PM | 323 14th St, Boston, MA 02215 | 12/24/2019 6:13:00 PM | 323 14th St | Boston | MA 02215 | 99.99 |
| 1 | 109.99 | 9/21/2019 5:45:00 PM | 708 Willow St, Boston, MA 02... | 12/21/2019 5:45:00 PM | 708 Willow St | Boston | MA 02215 | 109.99 |

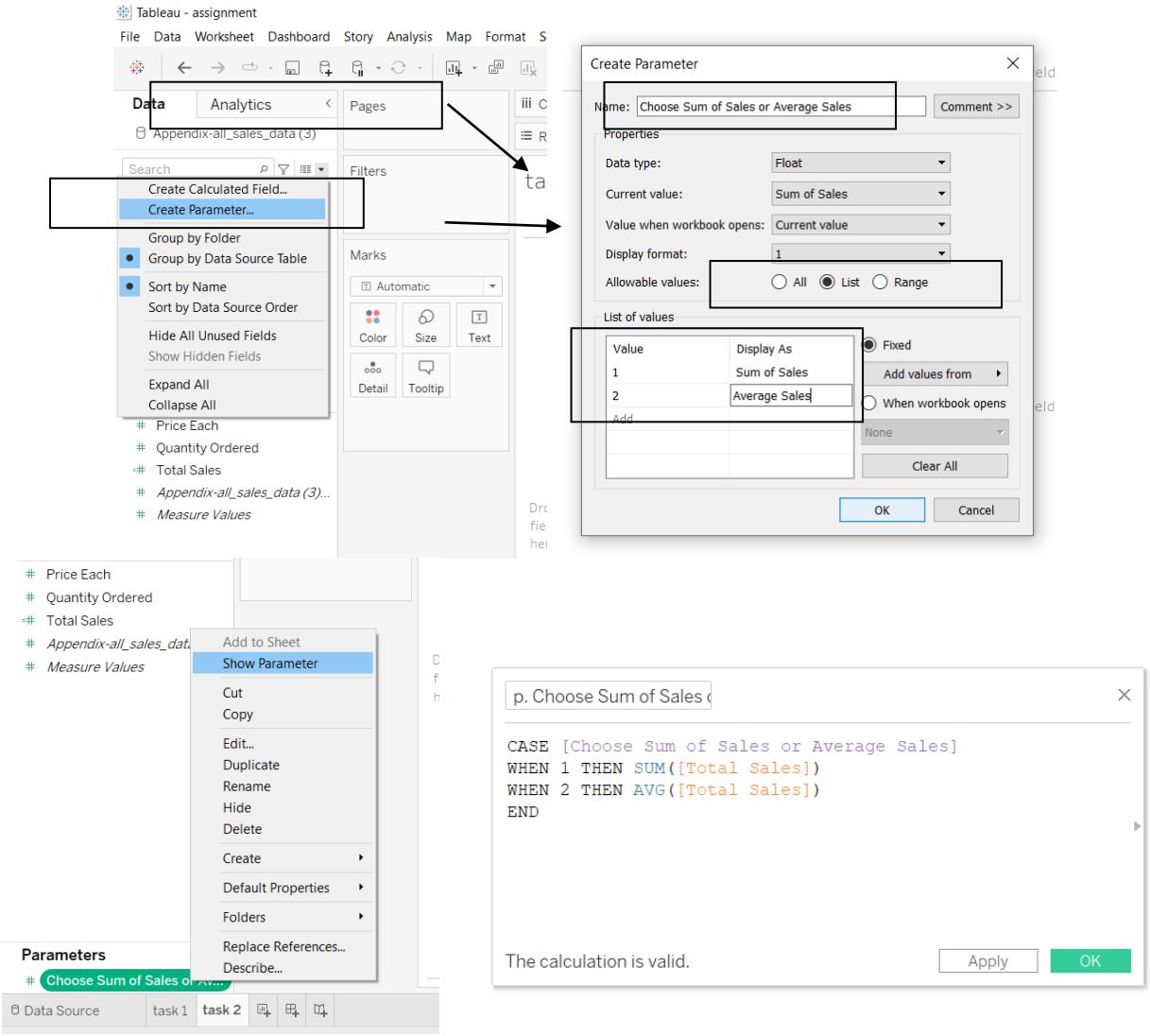
4. Displaying the total sales made in each City.



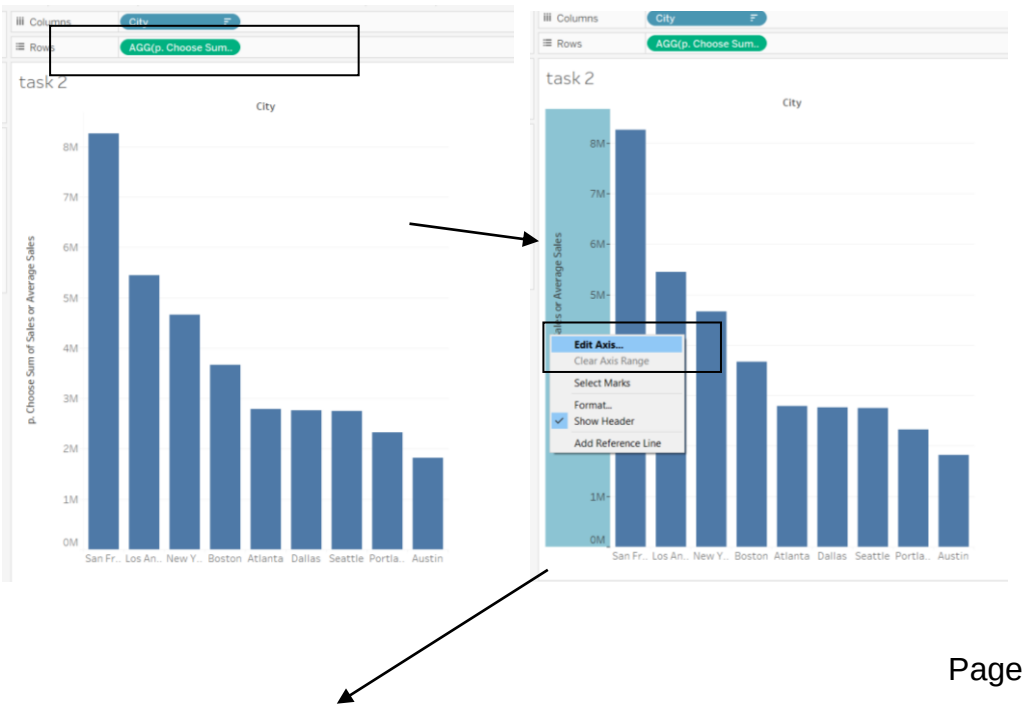

5. Ordering the result by the city with the highest sales by clicking sorted descending.

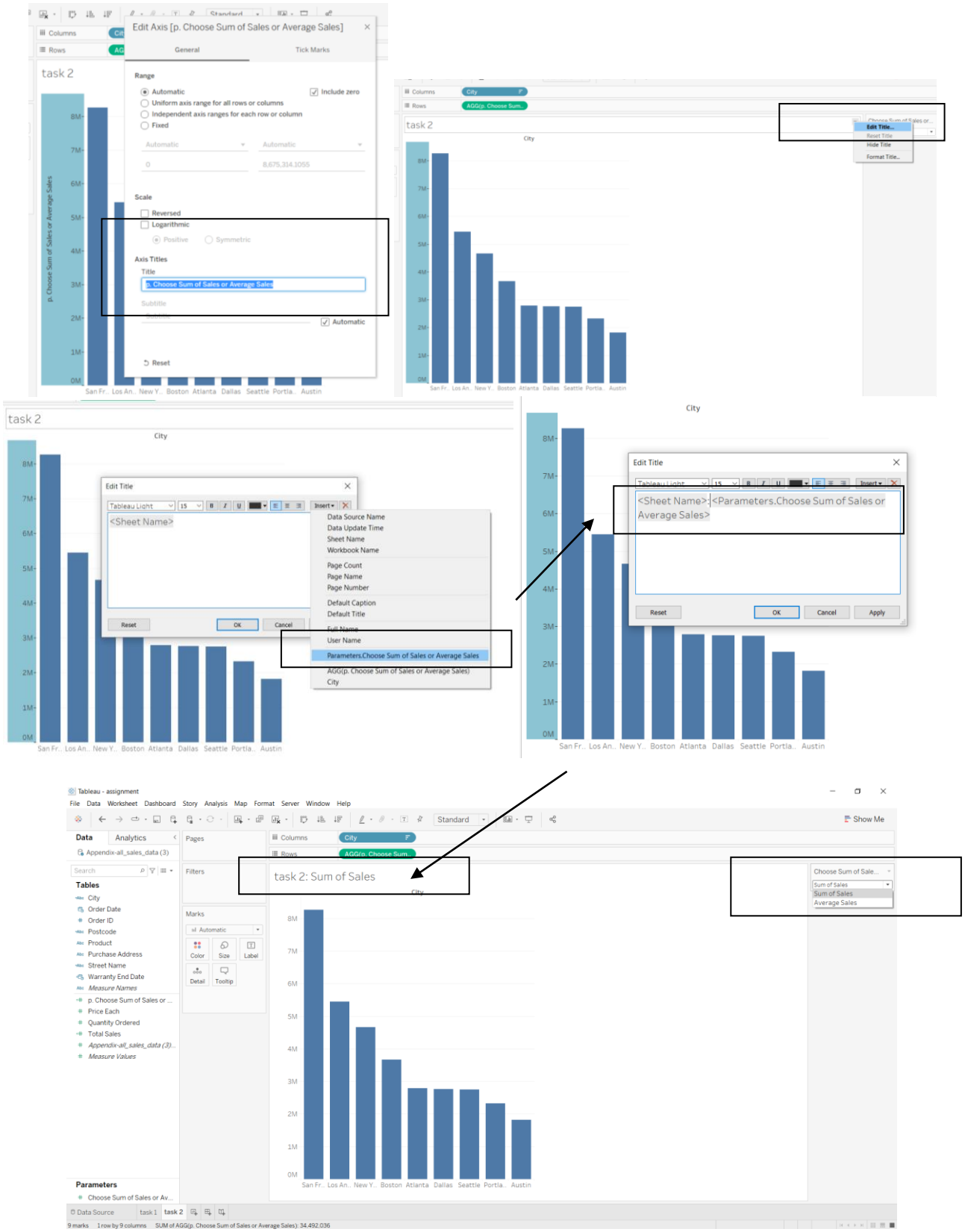


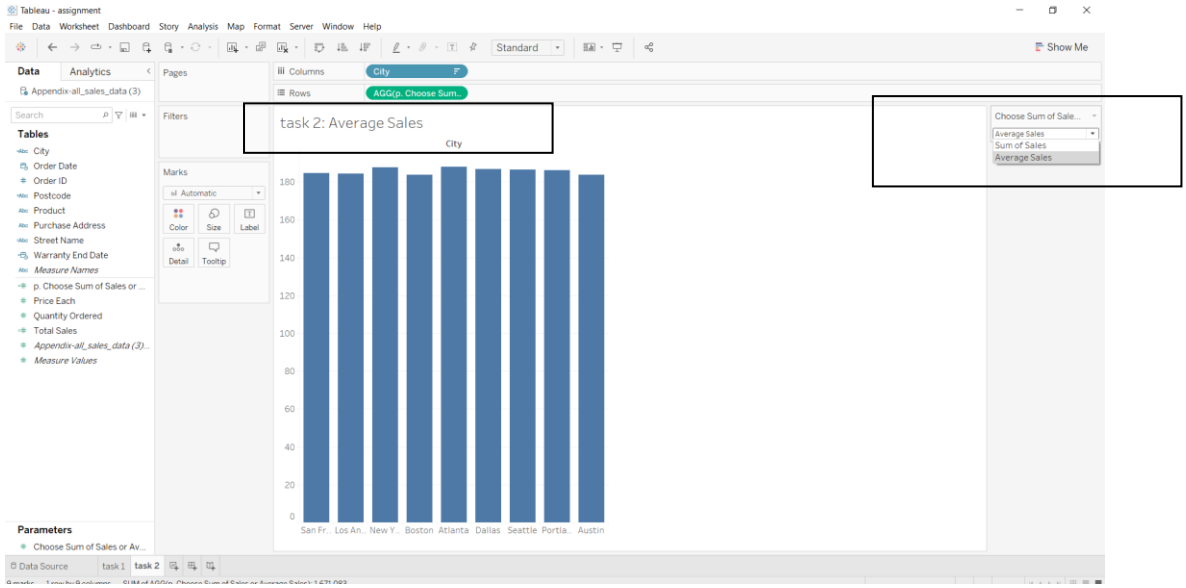

6. Creating a parameter to let user choose Sum of sales or Average Sales.




7. Giving the user an option to either choose Sum of Sales or Average Sales.



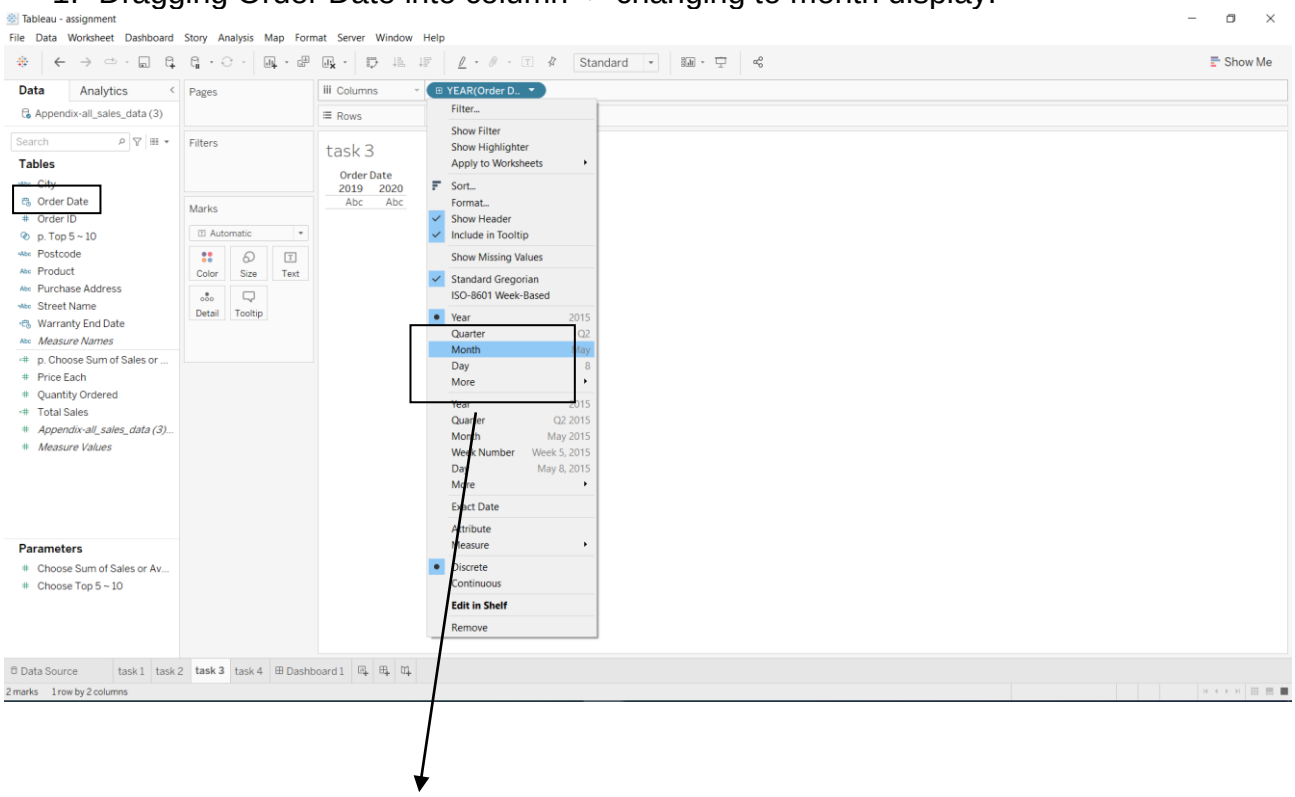

Edit Axis [p. Choose Sum of Sales or Average Sales]
General
Tick Marks
task 2
Range
Automatic
Uniform axis range for all rows or columns
Independent axis ranges for each row or column
Fixed
Include zero
Scale
Reversed
Logarithmic
Axis Titles
Title
p. Choose Sum of Sales or Average Sales
Subtitle
Automatic
Reset
City
Edit Title...
Reset Title
Hide Title
Format Title...
Edit Title
<Sheet Name>
<Parameters.Choose Sum of Sales or Average Sales>
Data Source Name
Data Update Time
Sheet Name
Workbook Name
Page Count
Page Name
Page Number
Default Caption
Default Title
User Name
Parameters.Choose Sum of Sales or Average Sales
AGG(p. Choose Sum of Sales or Average Sales)
City
Reset
OK
Cancel
Apply
Tableau - assignment
Data
Analytics
Tables
Parameters
task 2: Sum of Sales
Choose Sum of Sale...
Sum of Sales
Average Sales
San Fr..
Los An..
New Y..
Boston
Atlanta
Dallas
Seattle
Portla..
Austin

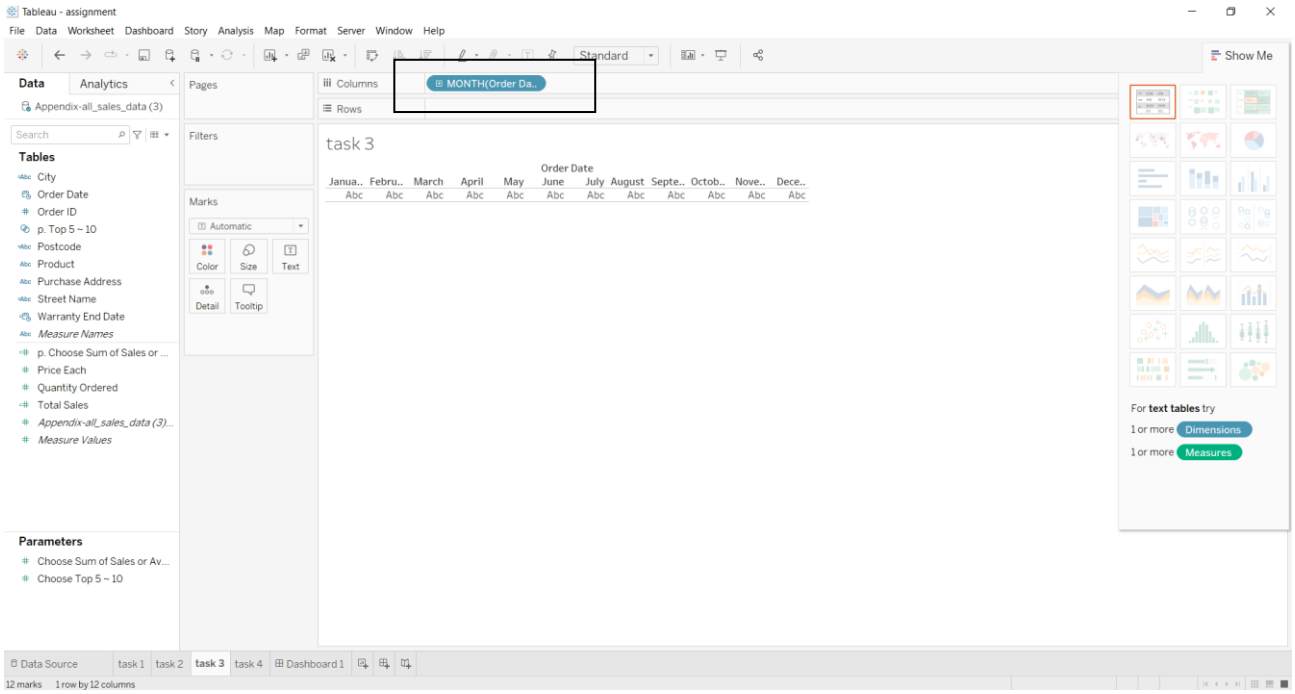



## 2.3 Task 3

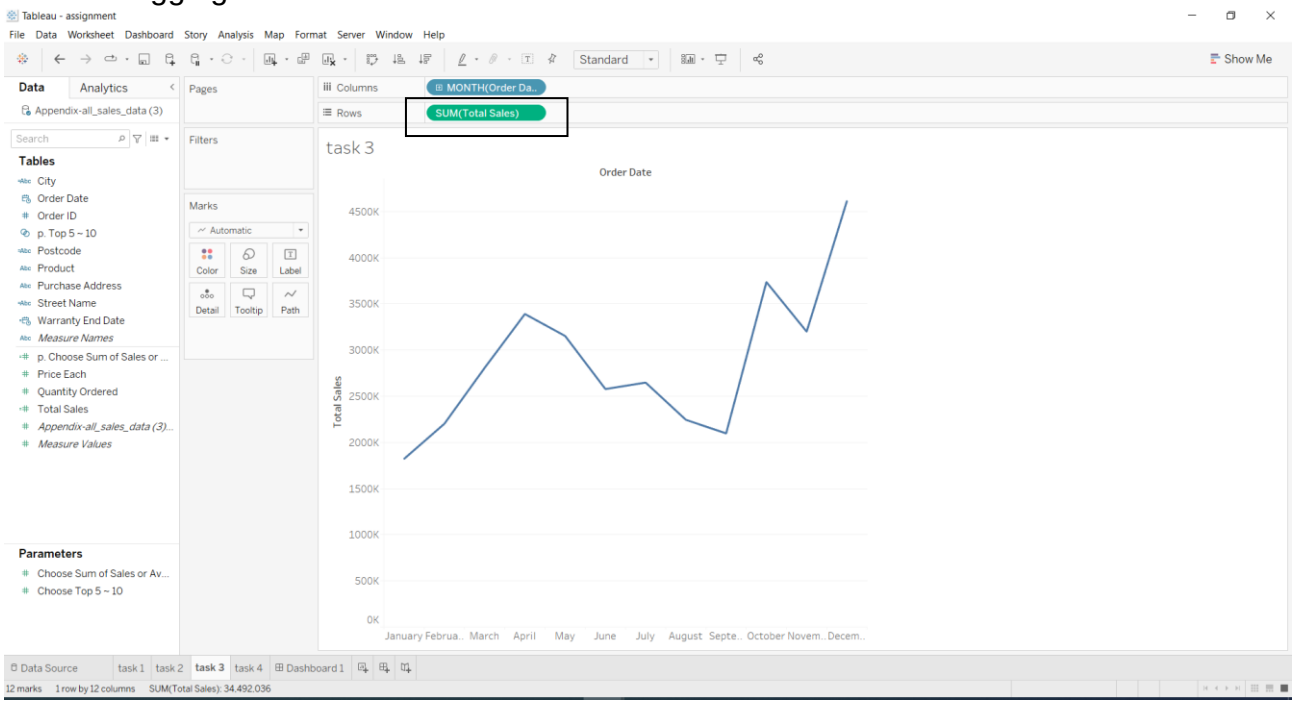
On a regular basis, providing performance of the stores and using an appropriate chart showing the average sales and total sales.

1. Dragging Order Date into column -> changing to month display.

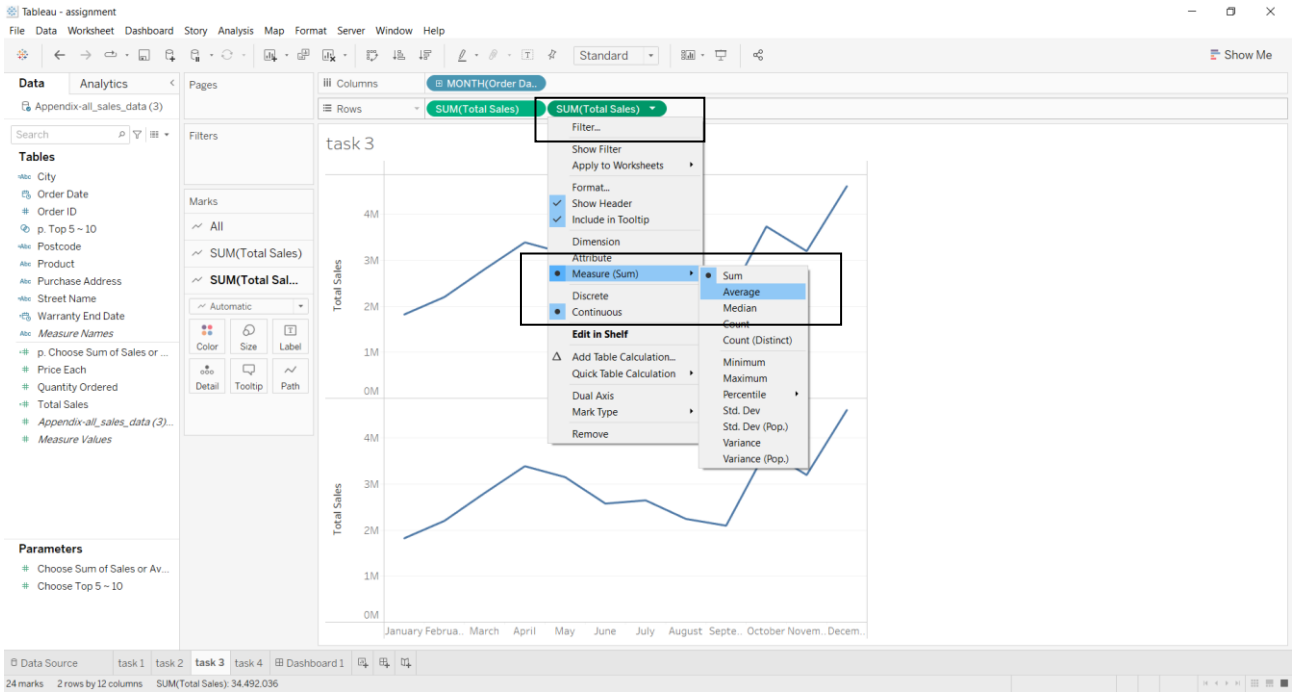




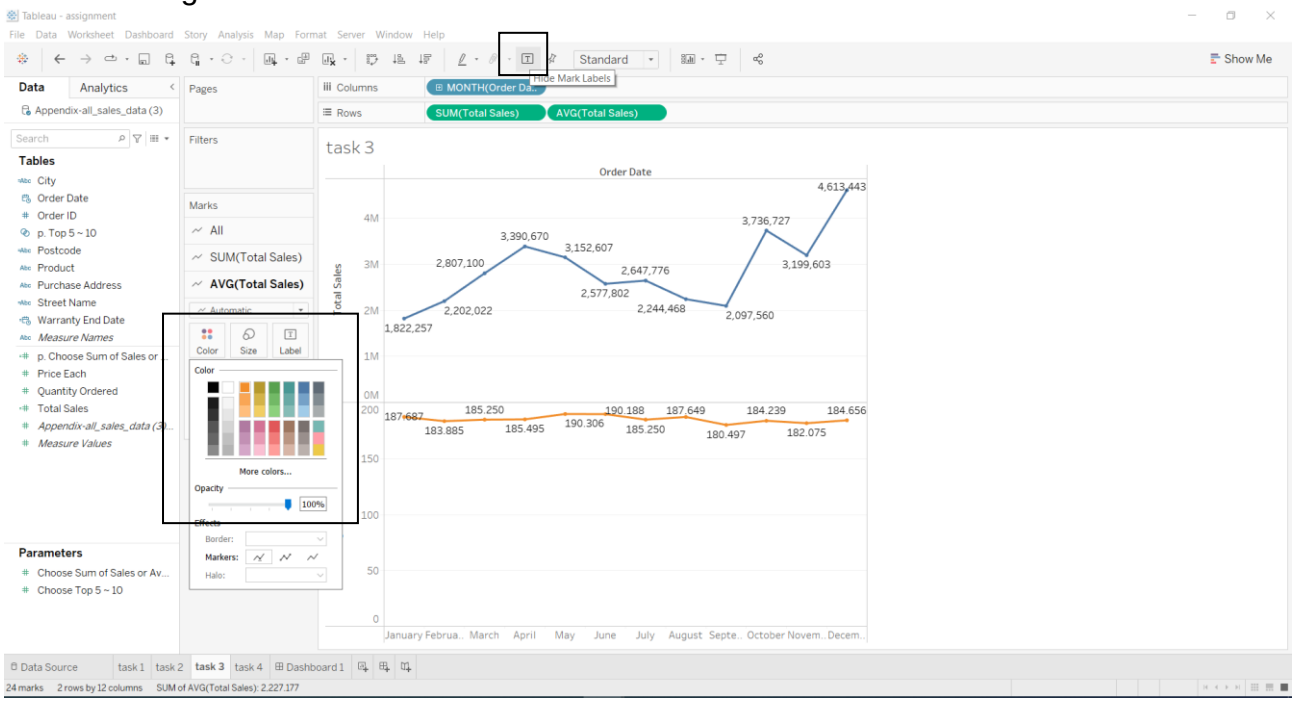

2. Dragging total sales into rows.

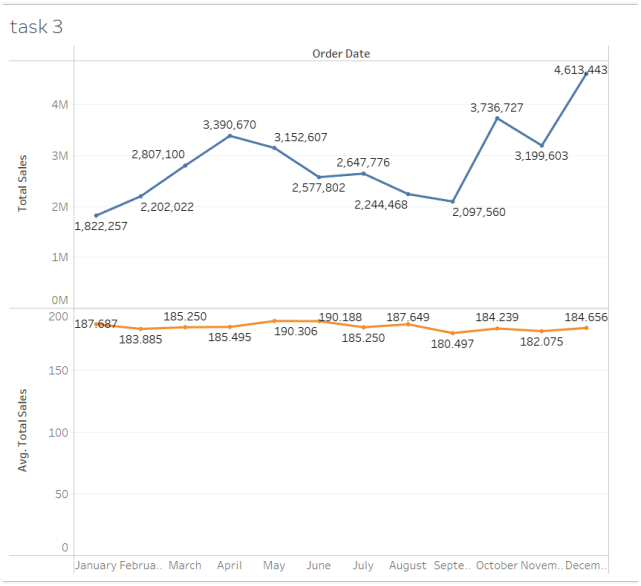



3. Dragging total sales into rows -> changing into AVG() function.




4. Adding labels and colors.

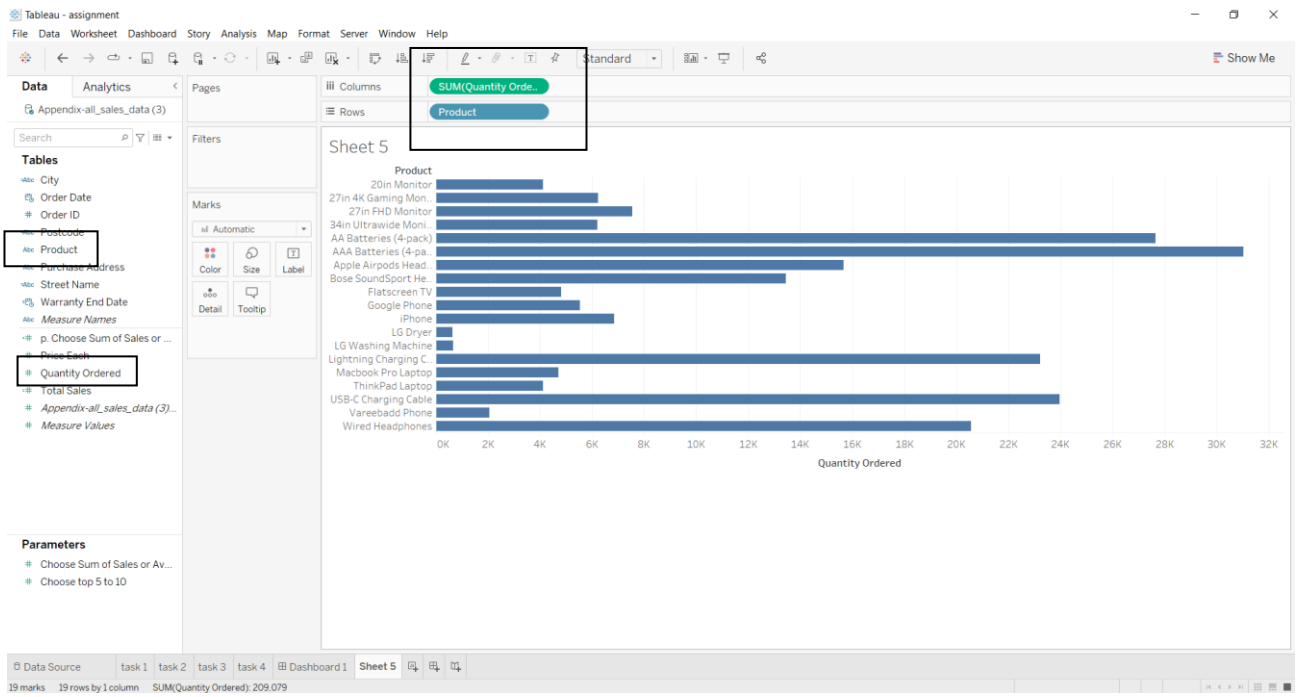




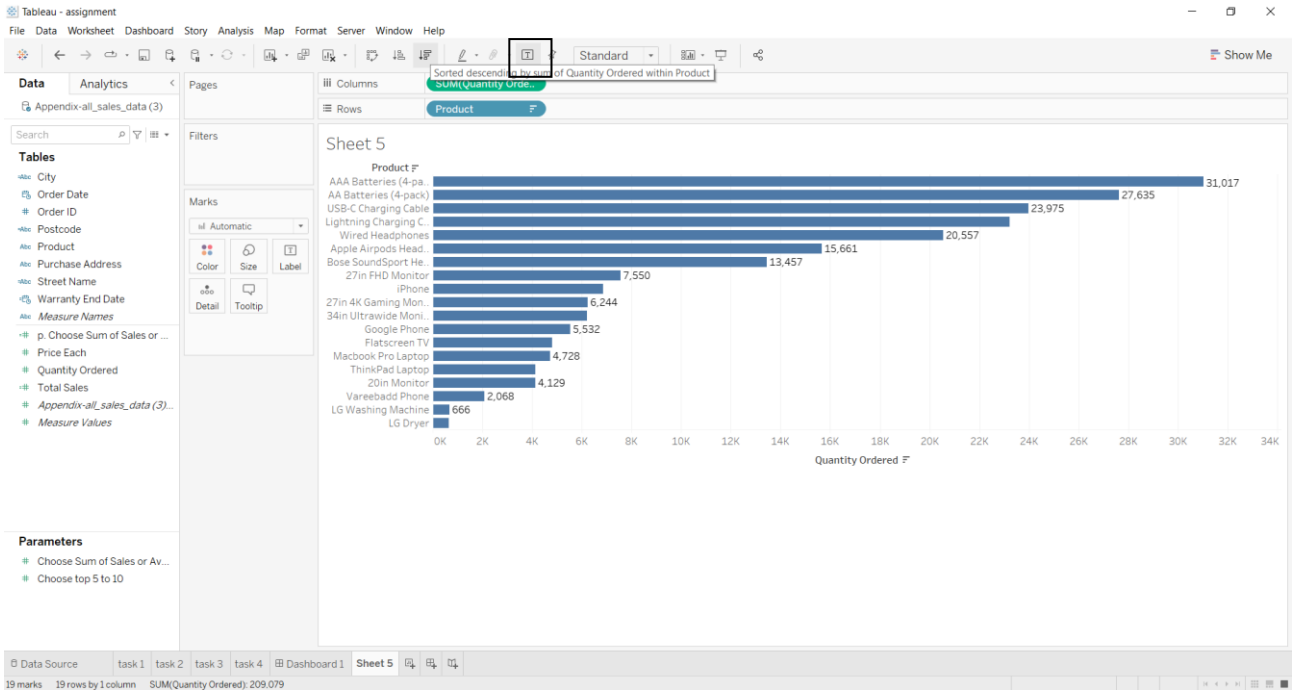

## 2.4 Task 4

Using an appropriate technique in Tableau, displaying a chart showing the top ten or top five selling product by using a parameter for management.

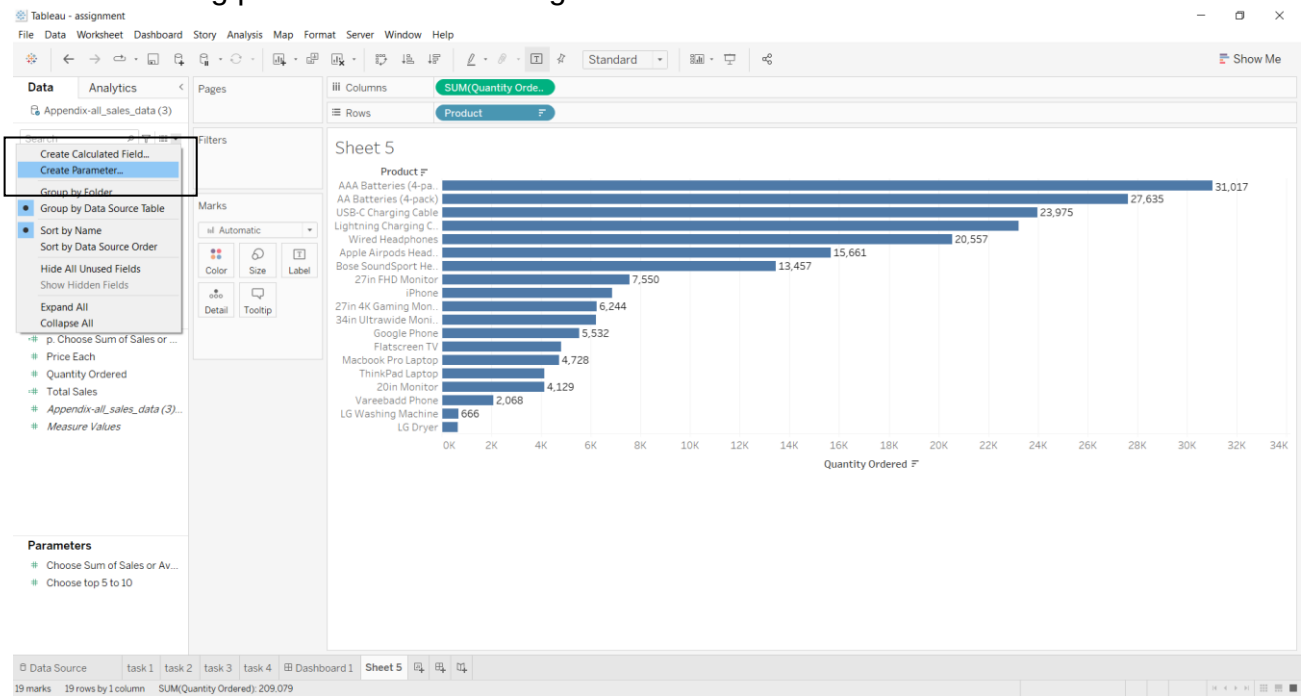
1. Dragging product and quantity ordered into rows and columns respectively.

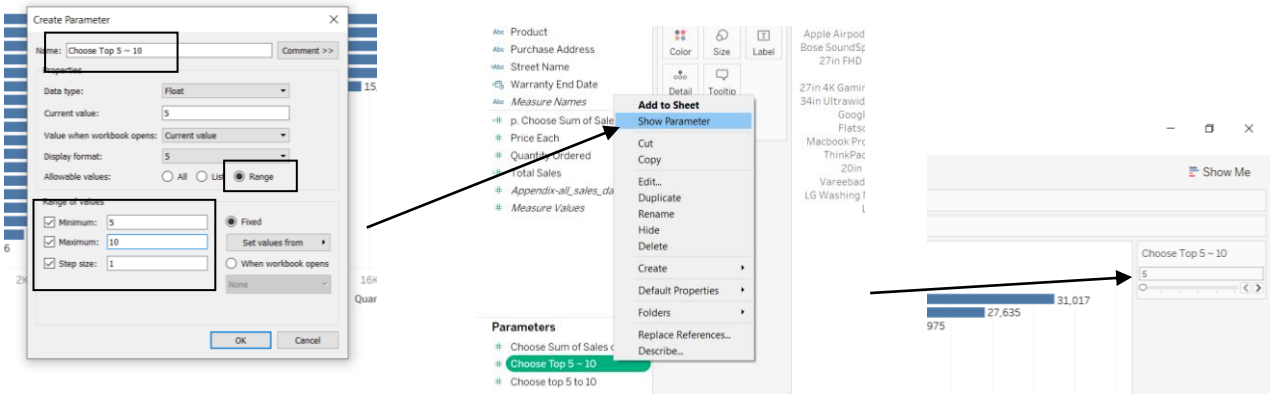



2. Clicking sorted descending by sum of Quanty within Product.

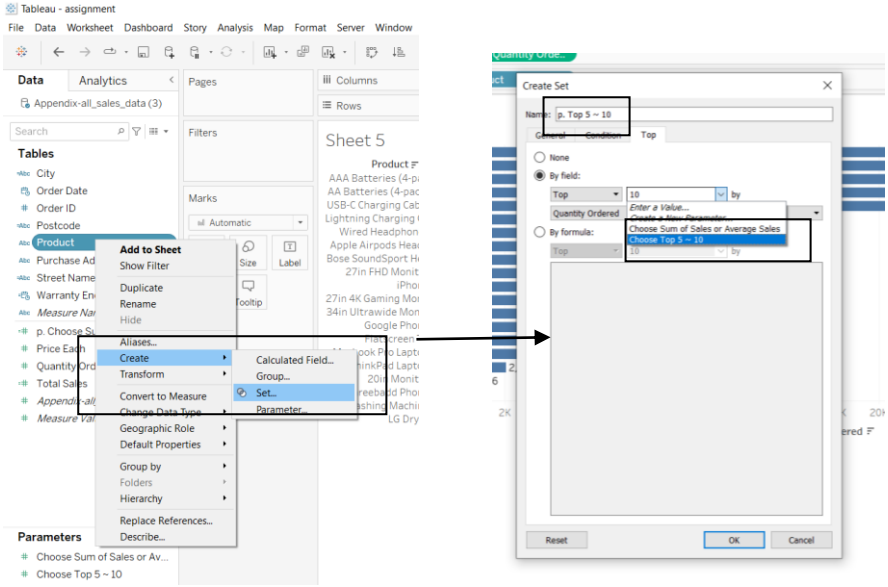



3. Creating parameter for choosing 5 to 10.

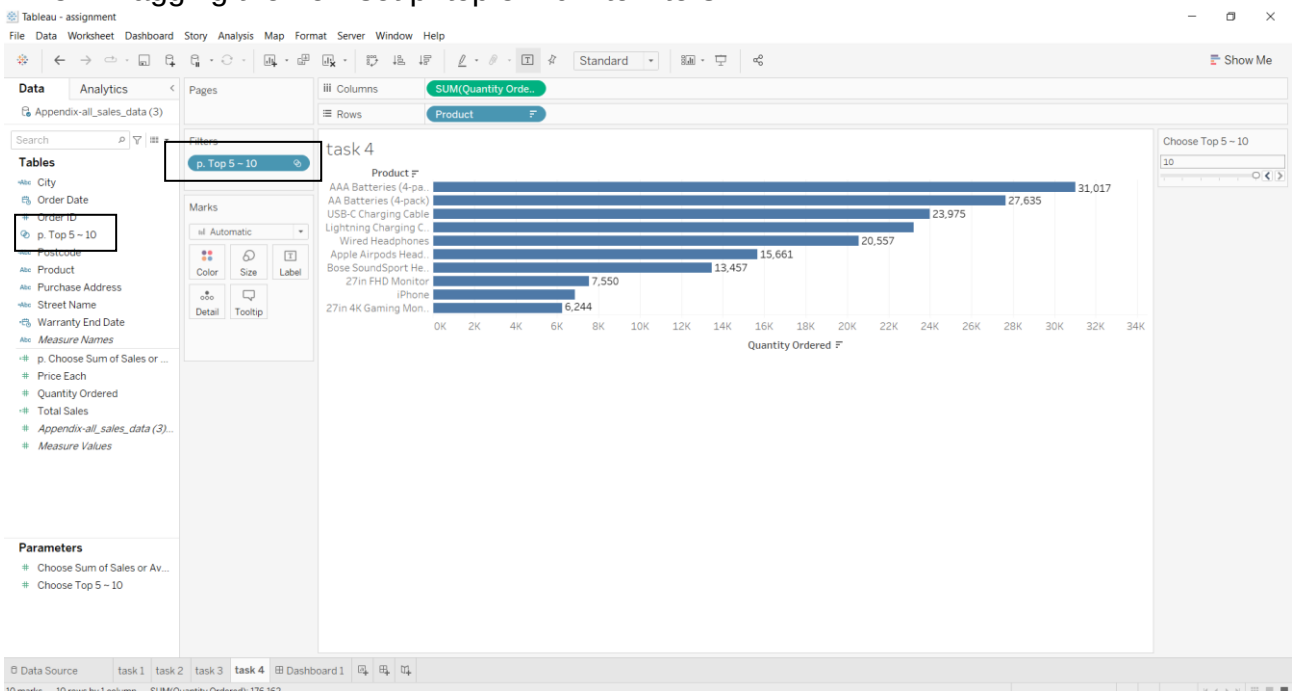




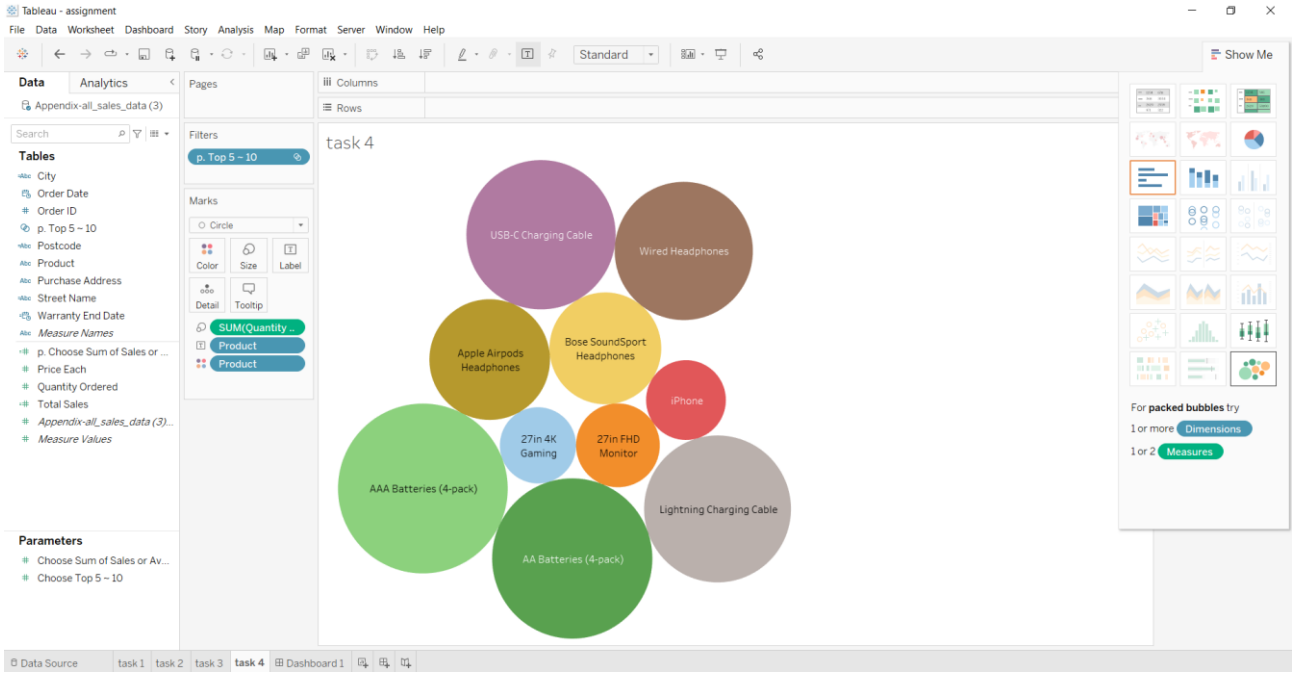

4. Create set on product to use the parameter.

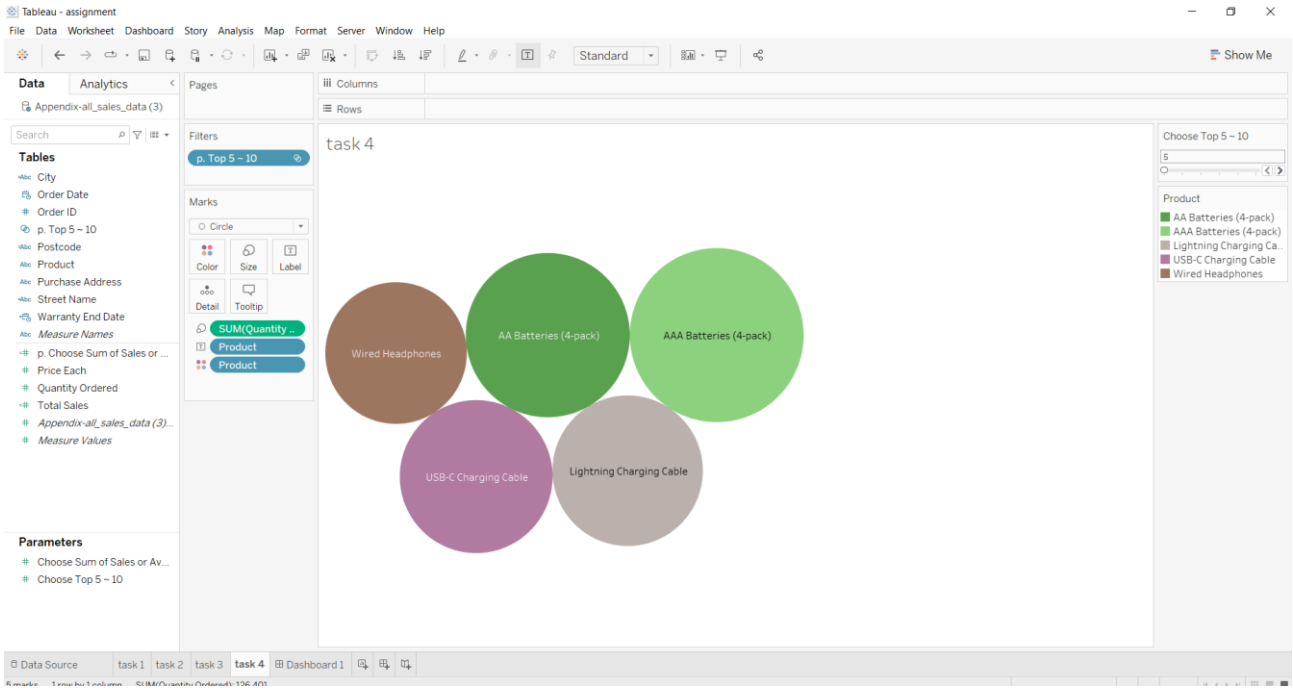



5. Dragging the new set p. top 5~10 into filters.




6.Using appropriate graph.

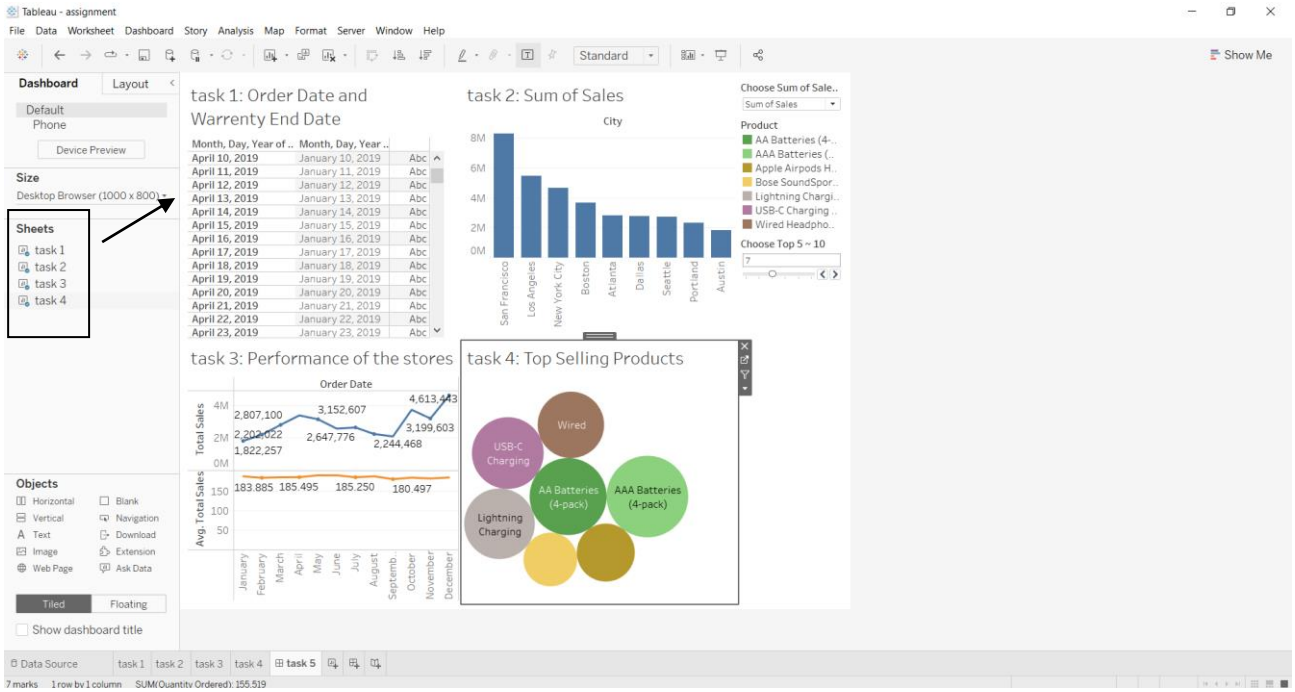


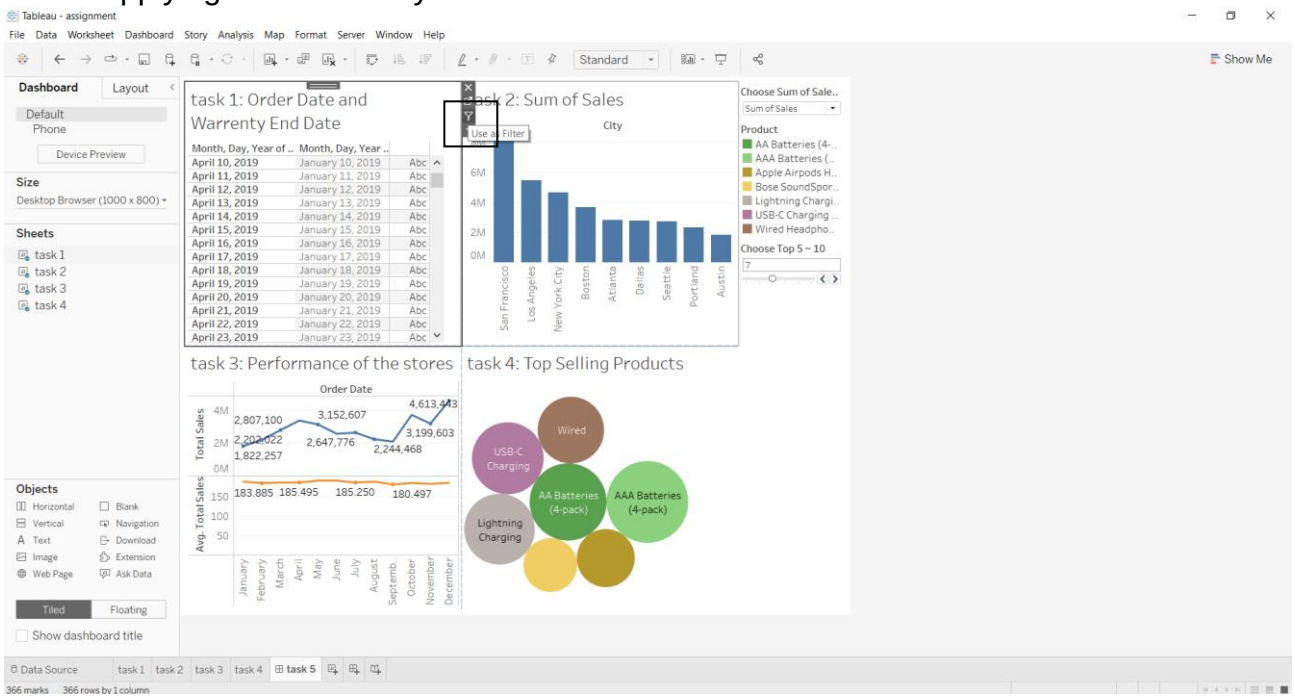

## 2.5 Task 5

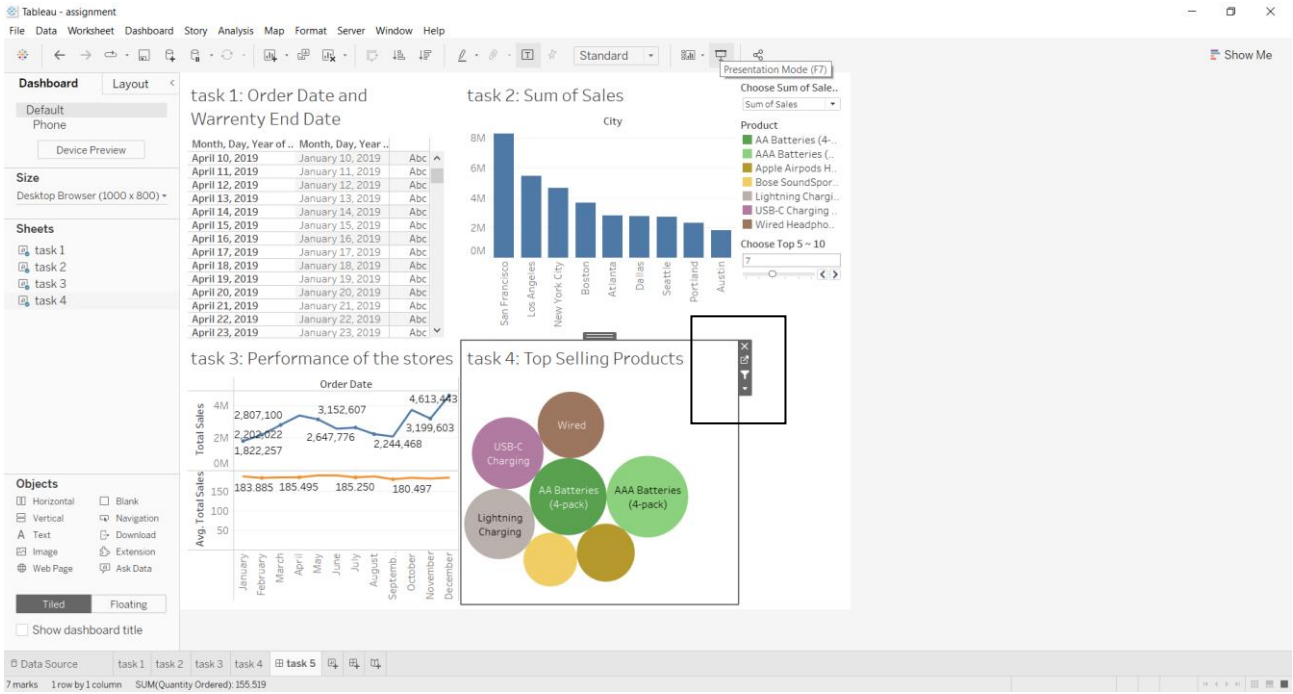
Creating an interactive dashboard with the previous four sheets.

1. Dragging the sheets into the dashboard.

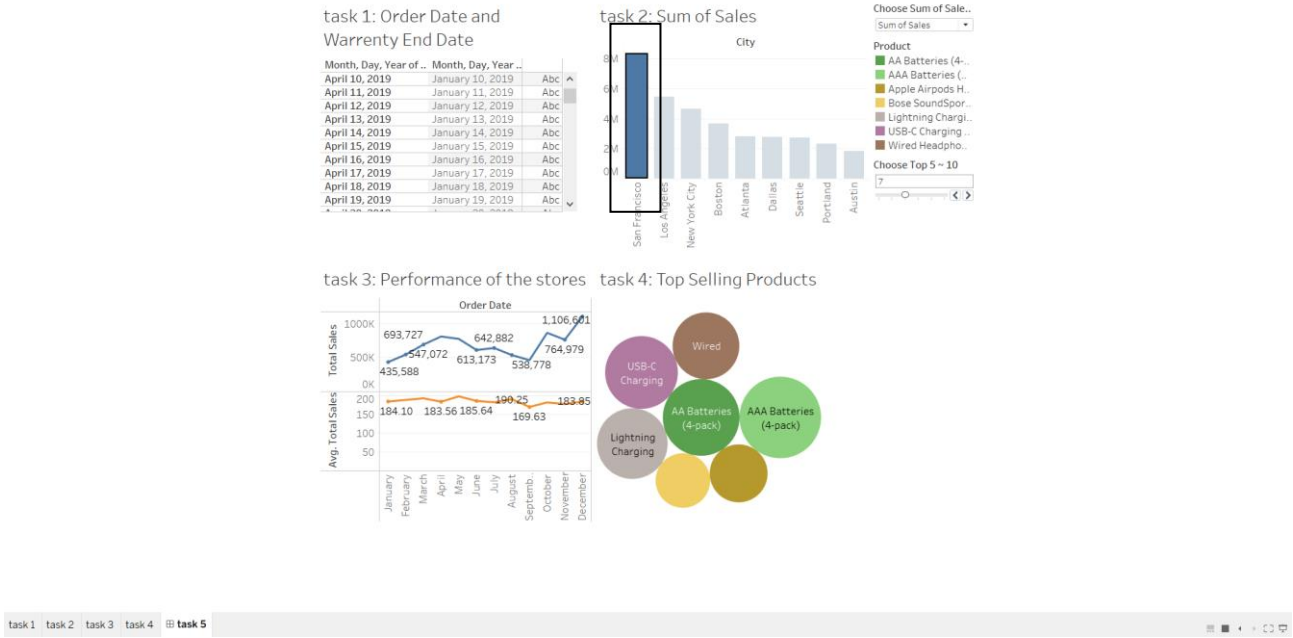


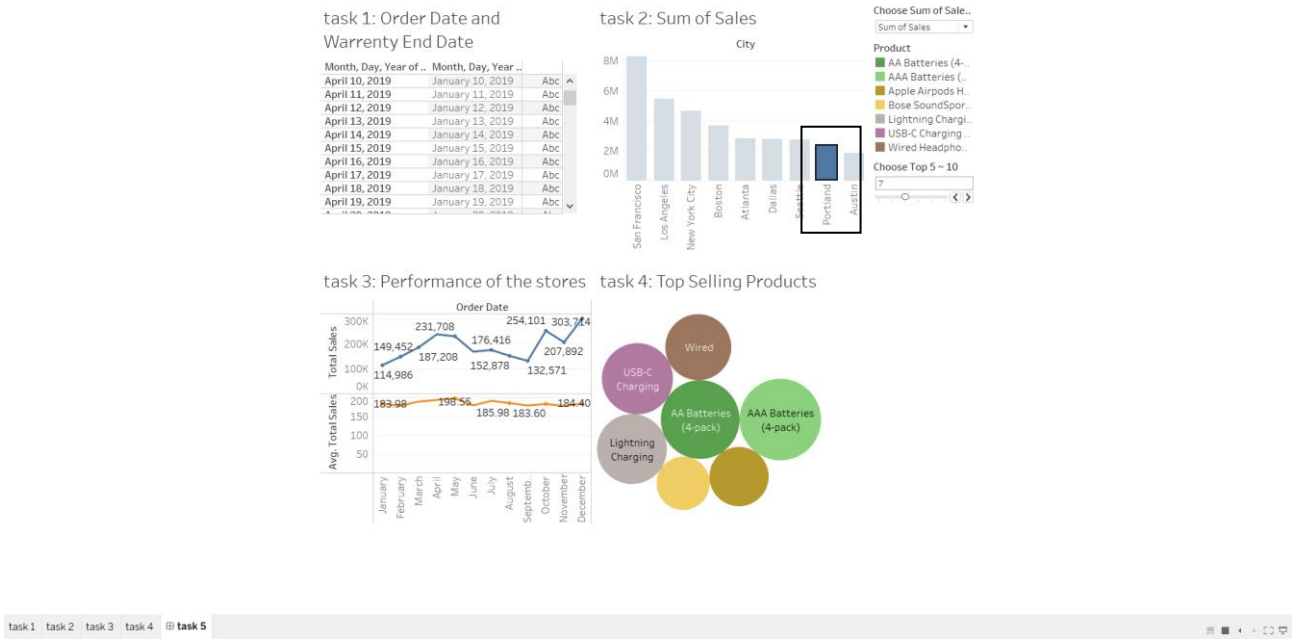

2. Applying filters to every sheets.

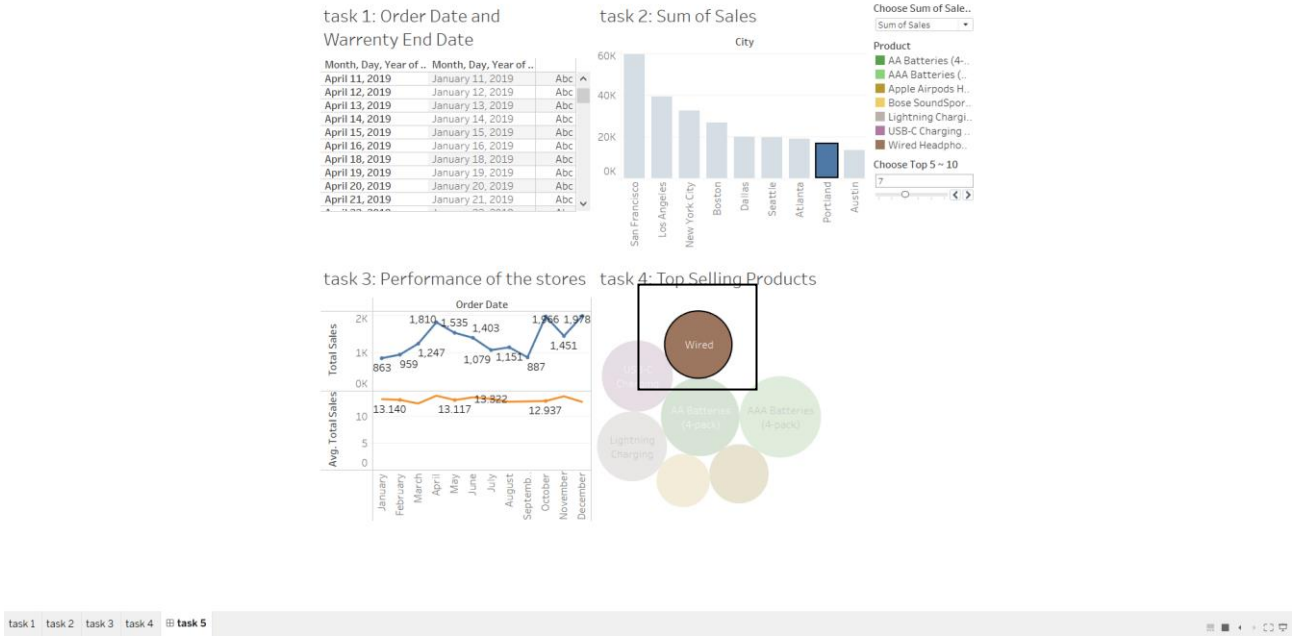


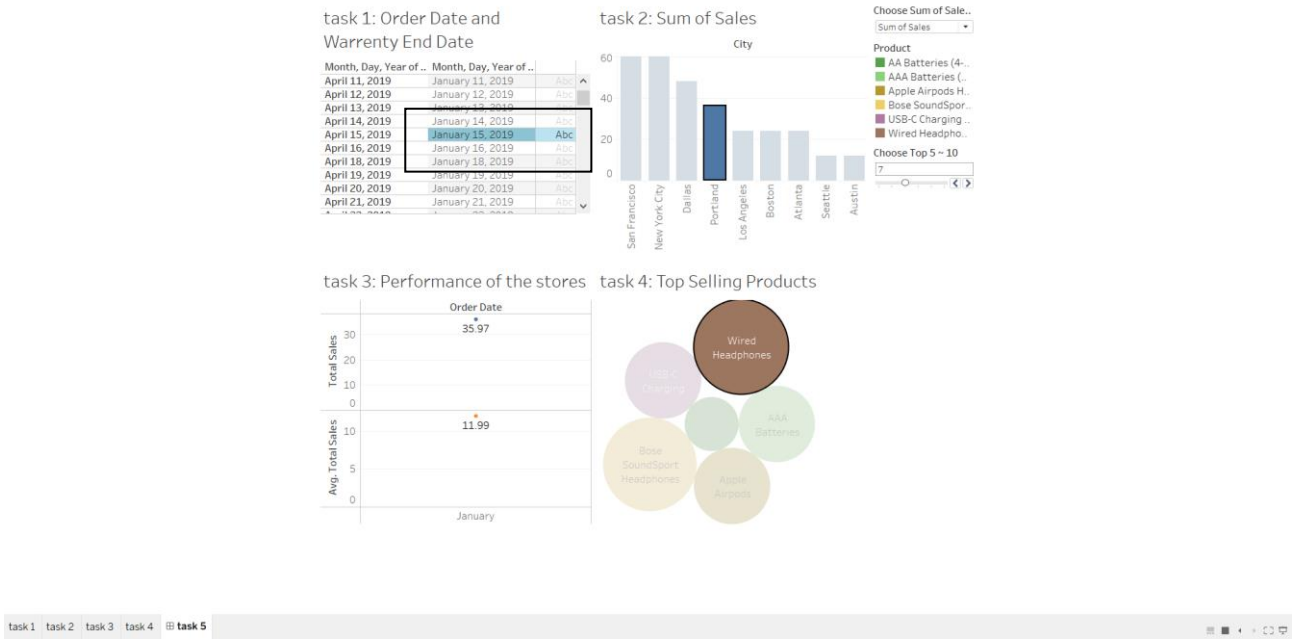



3. By clicking into one variable of one sheet, other are getting changed.




task 1: Order Date and Warrenty End Date
Month, Day, Year of .. Month, Day, Year ..
April 10, 2019 January 10, 2019 Abc
April 11, 2019 January 11, 2019 Abc
April 12, 2019 January 12, 2019 Abc
April 13, 2019 January 13, 2019 Abc
April 14, 2019 January 14, 2019 Abc
April 15, 2019 January 15, 2019 Abc
April 16, 2019 January 16, 2019 Abc
April 17, 2019 January 17, 2019 Abc
April 18, 2019 January 18, 2019 Abc
April 19, 2019 January 19, 2019 Abc
task 2: Sum of Sales
City
8M
6M
4M
2M
0M
San Francisco
Los Angeles
New York City
Boston
Atlanta
Dallas
Seattle
Portland
Austin
Choose Sum of Sale..
Sum of Sales
Product
AA Batteries (4-..
AAA Batteries (..
Apple Airpods H..
Bose SoundSpor..
Lightning Chargi..
USB-C Charging ..
Wired Headpho..
Choose Top 5 ~ 10
7
task 3: Performance of the stores
Order Date
Total Sales
300K
200K
100K
0K
149,452
231,708
187,208
114,986
176,416
152,878
254,101
132,571
303,774
207,892
Avg. Total Sales
200
150
100
50
183.98
198.55
185.98
183.60
184.40
January
February
March
April
May
June
July
August
Septemb.
October
November
December
task 4: Top Selling Products
Wired
USB-C Charging
AA Batteries (4-pack)
AAA Batteries (4-pack)
Lightning Charging
task 1
task 2
task 3
task 4
task 5


task 1: Order Date and Warrenty End Date
Month, Day, Year of .. Month, Day, Year of ..
April 11, 2019 January 11, 2019 Abc
April 12, 2019 January 12, 2019 Abc
April 13, 2019 January 13, 2019 Abc
April 14, 2019 January 14, 2019 Abc
April 15, 2019 January 15, 2019 Abc
April 16, 2019 January 16, 2019 Abc
April 18, 2019 January 18, 2019 Abc
April 19, 2019 January 19, 2019 Abc
April 20, 2019 January 20, 2019 Abc
April 21, 2019 January 21, 2019 Abc
task 2: Sum of Sales
City
60K
40K
20K
0K
San Francisco
Los Angeles
New York City
Boston
Dallas
Seattle
Atlanta
Portland
Austin
Choose Sum of Sale..
Sum of Sales
Product
AA Batteries (4-..
AAA Batteries (..
Apple Airpods H..
Bose SoundSpor..
Lightning Chargi..
USB-C Charging ..
Wired Headpho..
Choose Top 5 ~ 10
7
task 3: Performance of the stores
Order Date
Total Sales
2K
1K
0K
863
959
1,247
1,810
1,535
1,403
1,079
1,151
887
1,451
Avg. Total Sales
10
5
0
13,140
13,117
13,822
12,937
January
February
March
April
May
June
July
August
Septemb.
October
November
December
task 4: Top Selling Products
Wired
task 1
task 2
task 3
task 4
task 5




# Section 3 References

Gupta, B., 2022. *Analytics in HR - Data Driven Decision Making in Human Resources*. [online] Analytics India Magazine. Available at: <https://analyticsindiamag.com/analytics-in-hr-data-driven-decision-making-in-human-resources/> [Accessed 9 May 2022].

Jabir, B., Falih, N. and Rahmani, K., 2019. HR analytics a roadmap for decision making: case study. *Indonesian Journal of Electrical Engineering and Computer Science*, 15(2), p.979.
Kaggle.com. 2022. *HR Analytics*. [online] Available at: <https://www.kaggle.com/datasets/giripujar/hr-analytics> [Accessed 11 May 2022].

Kaggle.com. 2022. *Kaggle: Your Machine Learning and Data Science Community*. [online] Available at: <https://www.kaggle.com/> [Accessed 11 May 2022].

Opatha, H., 2020. HR Analytics: A Literature Review and New Conceptual Model. *International Journal of Scientific and Research Publications (IJSRP)*, 10(06), pp.130-141.

Rabhi, L., Falih, N., Afraites, A. and Bouikhalene, B., 2019. Big Data Approach and its applications in Various Fields: Review. *Procedia Computer Science*, 155, pp.599-605.

Roshini, S., Prakash, S., Shilpha Dharshini, J., Saroja, M. and Dhivya, J., 2021. Decision Tree and KNN Analysis for HR Analytics Data. *2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA)*,.

Setiawan, I., Suprihanto, S., Nugraha, A. and Hutahaean, J., 2020. HR analytics: Employee attrition analysis using logistic regression. *IOP Conference Series: Materials Science and Engineering*, 830(3), p.032001.

Technologies, B., 2022. *Big Data Technologies | Top 12 Technology of Big Data*. [online] EDUCBA. Available at: <https://www.educba.com/big-data-technologies/> [Accessed 11 May 2022].

Van Den Heuvel, S. and Bondarouk, T., 2016. "The Rise (and Fall?) of HR Analytics: The Future Application, Value, Structure, and System Support." *Academy of Management Proceedings*, 2016(1), p.10908.