Introduction to Data Management PROJECT REPORT

(Project Semester August-December 2021)

PROJECT REPORT

ON
SALES IN USA

submitted by

Shaik Malika Sulthana

11902389

programme : Bachelor of Technology

Section: K19BH

Course Code : INT217

Under the Guidance of

Ashu : 23631

Discipline of CSE/IT

Lovely School of Computer Science & Engineering
Lovely Professional University, Phagwara

DECLARATION

I, Shaik Malika Sulthana student of Computer Science & Engineering under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date:

Name of the student
Shaik Malika Sulthana
Registration No. 11902389
Signature

S. Malika

ACKNOWLEDGEMENT

Primarily I'd like to thank my mentor Ms. Ashu, whose valuable guidance has been the ones that helped me patch this project and make it full proof success in contribution towards the completion of this project. Last but not least I'd rather thanks to Lovely Professional University, and my parent's inspiration, who gave me this golden opportunity to learn many new things, to learn another aspects of life.

Shaik Malika Sulthana

CONTENTS:

INTRODUCTION

A database is a collection of related information stored in one or more computer files. Often the data is organized into tables in such a way that it can easily be updated, sorted, corrected, and filtered.

A simple database such as Excel holds all information about one subject in a single table. Relational databases, on the other hand, consist of many tables with each one containing information about different, but related topics.

- Data management is important because the data your organization creates is a very valuable resource.
- The last thing you want to do is spend time and resources collecting data and business intelligence, only to lose or misplace that information.
- In that case, you would then have to spend time and resources again to get that same business intelligence you already had.
- And on that data analysis is carried out which show visualization of our problems in efficient way.
- Data Analysis is a process of inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information, informing conclusions, and supporting decision- making.
- This project is based on such data analysis on IMDb data from 2006 to 2016
- IMDb is an online database of information related to films, television programs, home videos, video games, and streaming content online
- including cast, production crew and personal biographies, plot summaries, trivia, ratings, and fan and critical reviews.
- This IMDb dataset contains 12 data fields

OBJECTIVES/SCOPE OF ANALYSIS

After analysis of the dataset, the aim of this project is to give answer of given objectives in easy way:

- Total Revenue through each item (Pie- Doughnut)
- Revenue of each company (Bar – Clustered Bar)
- Sales by each employee in each year (column – Clustered column)
- Revenue in each region (Map – Filled Map)
- Sales trend in every month (Line – Line with a Marker)

SOURCE OF DATASET:

Source of dataset: **https://www.kaggle.com/datasets**

The dataset is based on 2000 Orders made from different companies in the years 2018 and 2019
The columns included in the dataset are given below:

- Order ID
  Id of each order

- Date
  The date of which order confirmed

- Company ID
  Id number of each company

- Company name
  Name of the company

- Sales Person
  Employee name who sales products

- Region
  Place where the sales happen

- Item
  Name of each item

- Price

  Price of each item

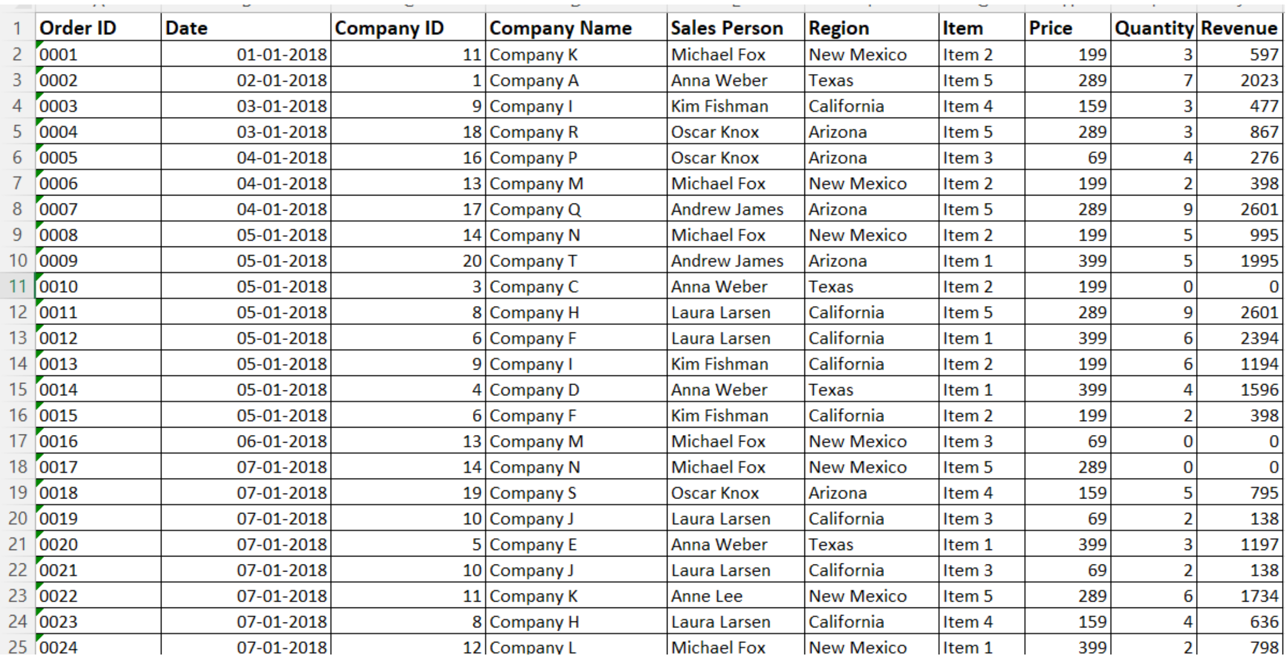
- Quantity

  Number of items brought

- Revenue

  Income through each order

Sample of dataset with data fields is given below:



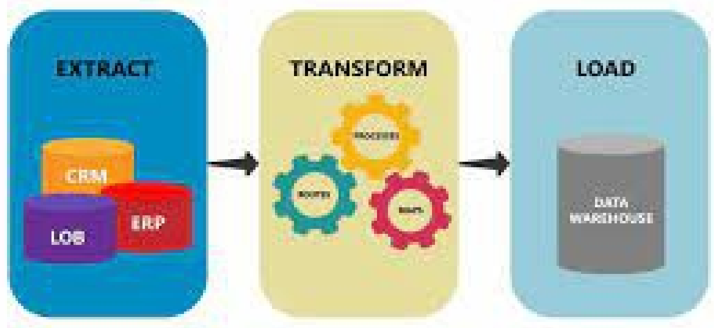
1	Order ID	Date	Company ID	Company Name	Sales Person	Region	Item	Price	Quantity	Revenue
2	0001	01-01-2018	11	Company K	Michael Fox	New Mexico	Item 2	199	3	597
3	0002	02-01-2018	1	Company A	Anna Weber	Texas	Item 5	289	7	2023
4	0003	03-01-2018	9	Company I	Kim Fishman	California	Item 4	159	3	477
5	0004	03-01-2018	18	Company R	Oscar Knox	Arizona	Item 5	289	3	867
6	0005	04-01-2018	16	Company P	Oscar Knox	Arizona	Item 3	69	4	276
7	0006	04-01-2018	13	Company M	Michael Fox	New Mexico	Item 2	199	2	398
8	0007	04-01-2018	17	Company Q	Andrew James	Arizona	Item 5	289	9	2601
9	0008	05-01-2018	14	Company N	Michael Fox	New Mexico	Item 2	199	5	995
10	0009	05-01-2018	20	Company T	Andrew James	Arizona	Item 1	399	5	1995
11	0010	05-01-2018	3	Company C	Anna Weber	Texas	Item 2	199	0	0
12	0011	05-01-2018	8	Company H	Laura Larsen	California	Item 5	289	9	2601
13	0012	05-01-2018	6	Company F	Laura Larsen	California	Item 1	399	6	2394
14	0013	05-01-2018	9	Company I	Kim Fishman	California	Item 2	199	6	1194
15	0014	05-01-2018	4	Company D	Anna Weber	Texas	Item 1	399	4	1596
16	0015	05-01-2018	6	Company F	Kim Fishman	California	Item 2	199	2	398
17	0016	06-01-2018	13	Company M	Michael Fox	New Mexico	Item 3	69	0	0
18	0017	07-01-2018	14	Company N	Michael Fox	New Mexico	Item 5	289	0	0
19	0018	07-01-2018	19	Company S	Oscar Knox	Arizona	Item 4	159	5	795
20	0019	07-01-2018	10	Company J	Laura Larsen	California	Item 3	69	2	138
21	0020	07-01-2018	5	Company E	Anna Weber	Texas	Item 1	399	3	1197
22	0021	07-01-2018	10	Company J	Laura Larsen	California	Item 3	69	2	138
23	0022	07-01-2018	11	Company K	Anne Lee	New Mexico	Item 5	289	6	1734
24	0023	07-01-2018	8	Company H	Laura Larsen	California	Item 4	159	4	636
25	0024	07-01-2018	12	Company L	Michael Fox	New Mexico	Item 1	399	2	798

ETL PROCESS:

In most organizations, data goes through an ETL (extract, transform and load) process before it is available for reporting. During the ETL process, data is extracted from a data source, then transformed, validated, standardized, corrected, quality checked and ultimately loaded into a data repository—such as a data mart or data warehouse—where it is streamlined for analysis and reporting.
Full form of ETL is Extract, Transform and Load.

The triple combination of ETL provides crucial functions that are many times combined into a single application or suite of tools that help in the following areas:

- Enhances Business Intelligence solutions for decision making.
- Allows verification of data transformation, aggregation and calculations rules.
- Allows sample data comparison between source and target system.
- Helps to improve productivity as it codifies and reuses without additional technical skills.

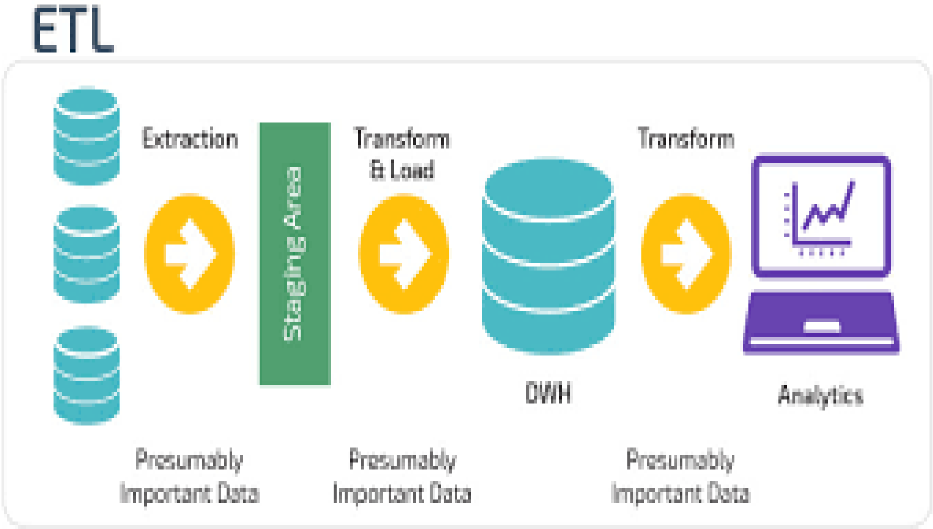


ETL - Extract, Transform, Load

ETL Process Implementation: Three Easy Steps

The acronym E-T-L can be divided into three phases which implement the entire process.

1. **E** – Extraction
2. **T** – Transformation
3. **L** – Loading



ETL
Extraction
Staging Area
Transform & Load
Transform
DWH
Analytics
Presumably Important Data
Presumably Important Data
Presumably Important Data

1st Step – Extraction

Before you can begin organizing your data, the first step in the ETL data process is to pull or extract the data from all the relevant sources and compile it. This ETL requirement and gathering process will include the necessary preparation for carrying out data integration. The data sources may include data from multiple sources: on-premise databases, CRM systems, marketing automation platforms, cloud data warehouses, unstructured and structured files, cloud applications, and any other sources you wish to draw insights from via analytical processing.

Once all the critical data has been consolidated, you'll notice that data from different sources is dated and structured in different formats. In this step, the compiled data must be organized according to date, size, and source to suit the transformation process. A certain level of consistency is required in all the data to be fed into the system and converted in the next step. The complexity of this step can vary significantly, depending on data types, the volume of data, and data sources.

ETL Extraction Steps

- Compile data from relevant sources.
- Organize data to make it consistent.

2nd Step – Transformation

Data Transformation is the second step of the ETL process in data integrations. In the first step, the ETL deployment was carried out. Now, in the second ETL phase, the ETL transformation is carried out: data extracted from the sources are compiled, converted, reformatted, and cleansed in the staging area to be fed into the target database in the next step.

The transformation step involves executing a series of functions and applying sets of rules to the extracted data to convert it into a standard format to meet the schema requirements of the target database. The level of manipulation required in ETL transformation depends solely on the data extracted and the needs of the business. It includes validation of data as well as rejection if they're not acceptable.

Quality data sources won't require many transformations, while other datasets might require it significantly. To meet your target database's technical and business requirements, you can subject it to several transformation techniques.

Following are Data Integrity Problems:

1. Different spelling of the same person like Jon, John, etc.
2. There are multiple ways to denote company name like Google, Google Inc.
3. Use of different names like Cleaveland, Cleveland.
4. There may be a case that different account numbers are generated by various applications for the same customer.
5. In some data required files remains blank
6. Invalid product collected at POS as manual entry can lead to mistakes.

Validations are done during this stage

- Filtering – Select only certain columns to load
- Using rules and lookup tables for Data standardization
- Character Set Conversion and encoding handling
- Conversion of Units of Measurements like Date Time Conversion, currency conversions, numerical conversions, etc.

•Data threshold validation check. For example, age cannot be more than two digits.
•Data flow validation from the staging area to the intermediate tables.
•Required fields should not be left blank.
•Cleaning (for example, mapping NULL to 0 or Gender Male to “M” and Female to “F” etc.)
•Split a column into multiples and merging multiple columns into a single column.
•Transposing rows and columns,
•Use lookups to merge data
•Using any complex data validation (e.g., if the first two columns in a row are empty then it automatically reject the row from processing)

3rd Step – Loading

The concluding step in the three-step data ETL process is loading the datasets that have been extracted and transformed earlier into the target database. There are two ways to go about it; the first is a SQL insert routine that involves the manual insertion of each record in every row of your target database table. The other loading approach uses a bulk load of data, reserved for massive data loading.

The SQL insert may be slow, but it conducts data quality checks with each entry. While the bulk load is much faster for loading massive amounts of data, it does not consider data integrity for every record. Bulk loading is ideal for datasets you’re confident are free of errors.

Types of Loading:

•**Initial Load** — populating all the Data Warehouse tables
•**Incremental Load** — applying ongoing changes as when needed periodically.
•**Full Refresh** —erasing the contents of one or more tables and reloading with fresh data.

Load verification

•Ensure that the key field data is neither missing nor null.
•Test modeling views based on the target tables.
•Check that combined values and calculated measures.
•Data checks in dimension table as well as history table.
•Check the BI reports on the loaded fact and dimension table.

Analysis on dataset

1.Total Revenue through each item

- **Introduction**

  By performing this analysis, we will get know the total Revenue that the company gets by selling each item.

- **Description**

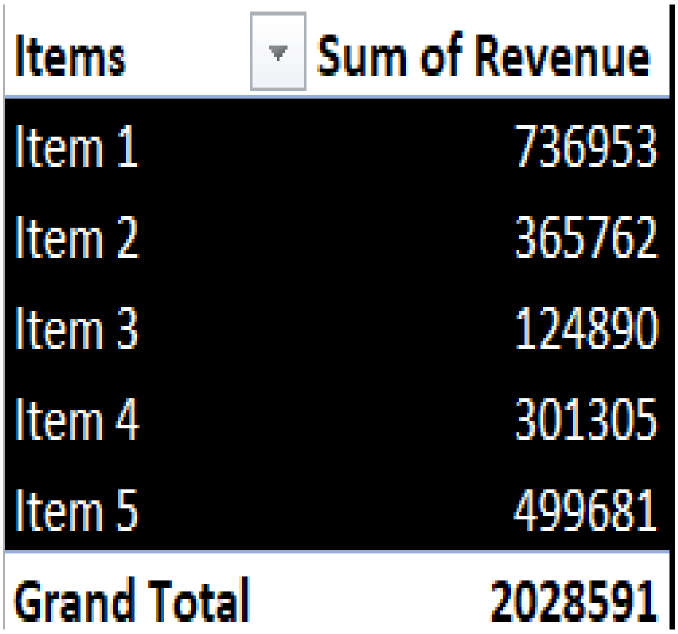
  The analysis is based on Item , Revenue

- **Specific requirements, functions and formulas**

  Pivot table is used for the analysis.

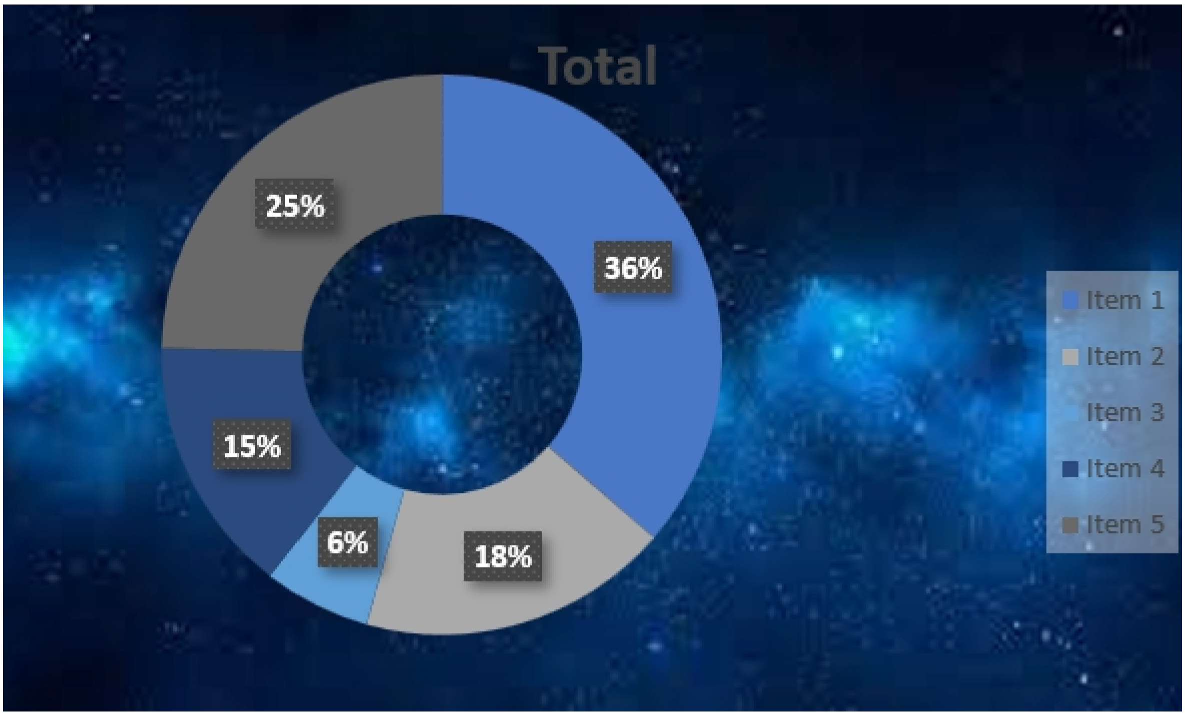
  Sum function is used in pivot table for the sum of the revenues in the pivot table

- **Analysis results**



Items	Sum of Revenue
Item 1	736953
Item 2	365762
Item 3	124890
Item 4	301305
Item 5	499681
Grand Total	**2028591**

- **Visualization**



2. Revenue of each company

- **Introduction**

  By performing this analysis, we will get Revenue gained by the each company by selling Items.

- **Description**

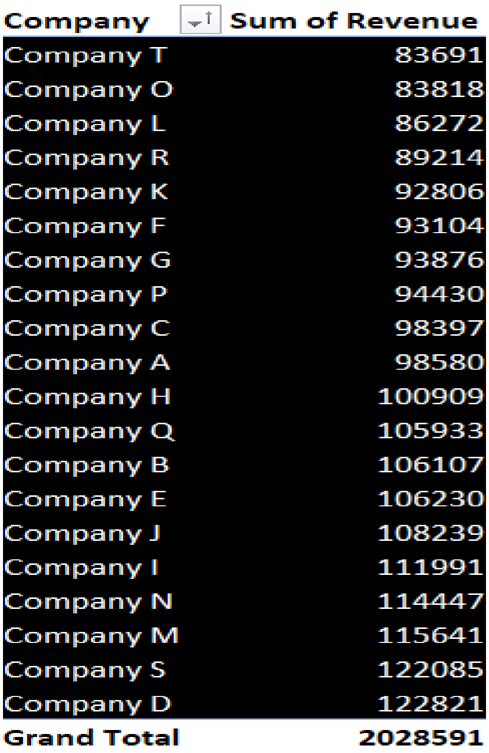
  The analysis based on the Company, Revenue.

- **Specific requirements, functions and formulas**

  Pivot table is used for the analysis.

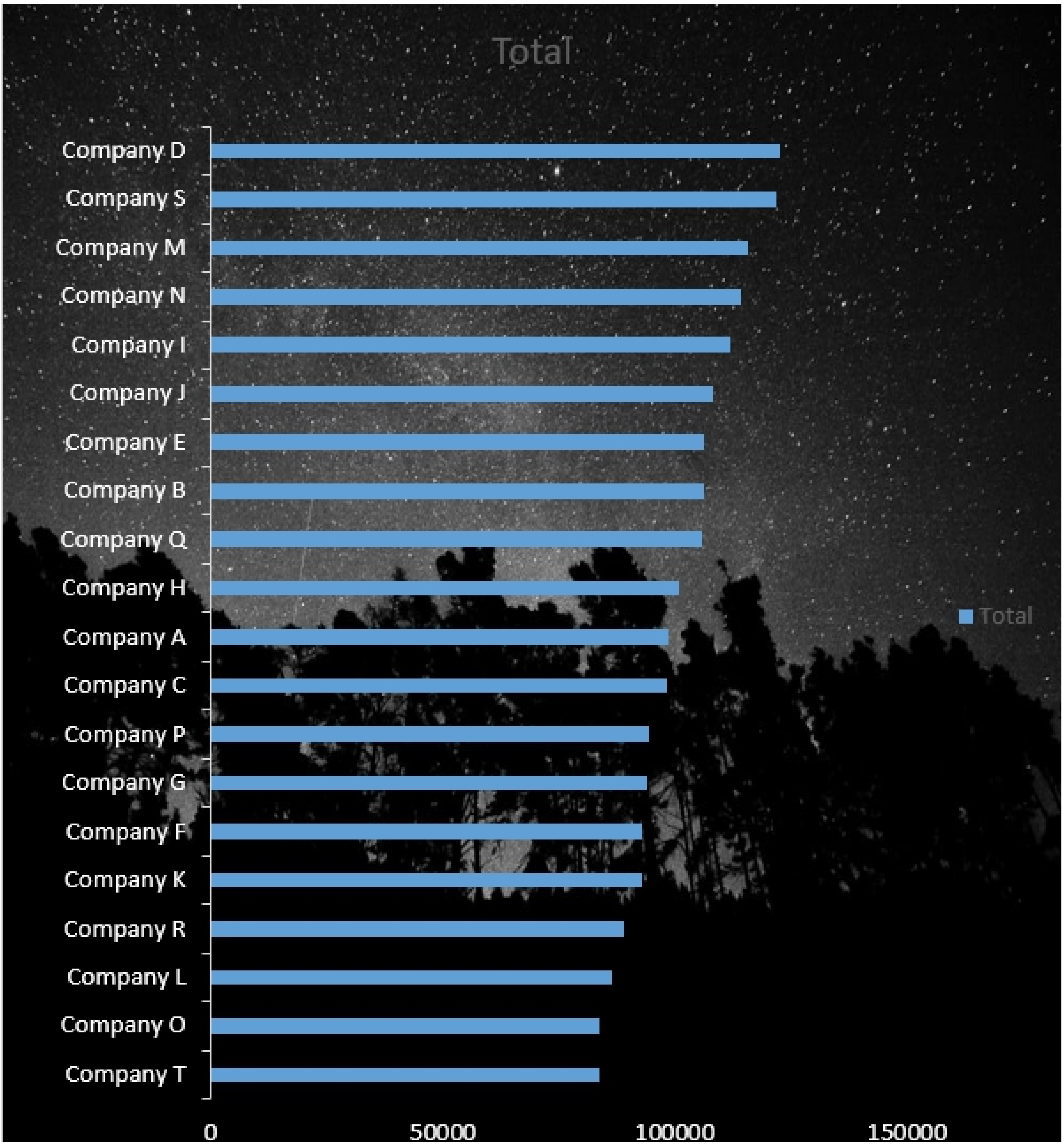
  Sum function is used in pivot table for the sum of the revenues in the pivot table

- **Analysis results**



Company	Sum of Revenue
Company T	83691
Company O	83818
Company L	86272
Company R	89214
Company K	92806
Company F	93104
Company G	93876
Company P	94430
Company C	98397
Company A	98580
Company H	100909
Company Q	105933
Company B	106107
Company E	106230
Company J	108239
Company I	111991
Company N	114447
Company M	115641
Company S	122085
Company D	122821
Grand Total	**2028591**

- **Visualization**



3. Sales by each employee in each year

- **Introduction**

  By performing this analysis, we will get Revenue gained by each Employee by selling items in each year.

- **Description**

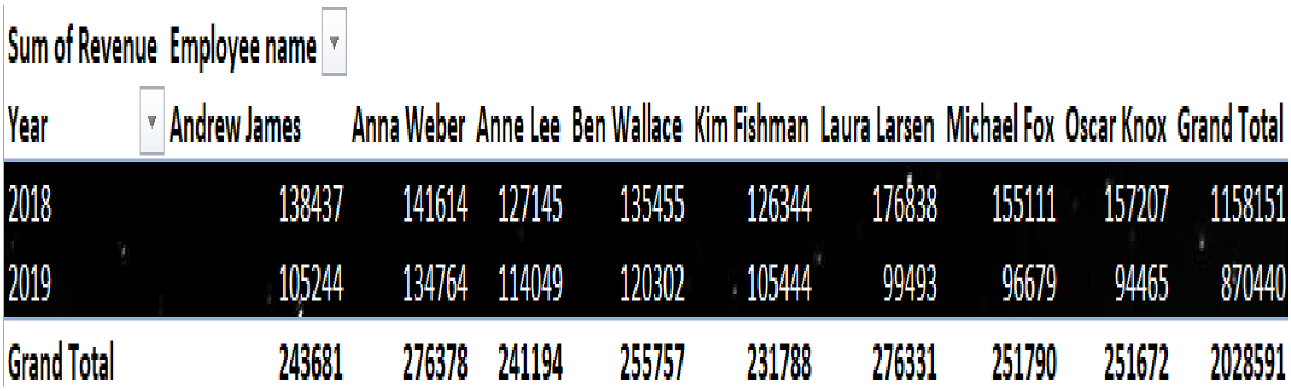
  The analysis based on Year and Sales Person.

- **Specific requirements, functions and formulas**

  Pivot table is used for the analysis.

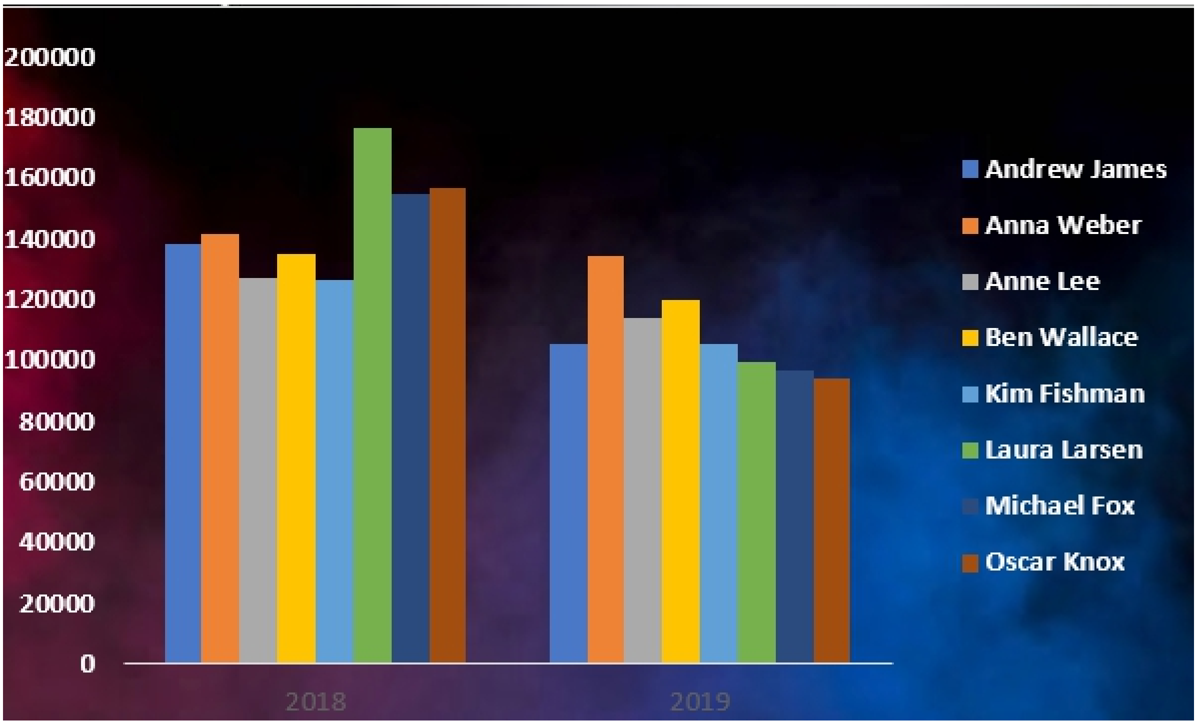
  Sum function is used in pivot table for the sum of the revenues in the pivot table

- **Analysis results**



Sum of Revenue	Employee name ▾								
Year ▾	Andrew James	Anna Weber	Anne Lee	Ben Wallace	Kim Fishman	Laura Larsen	Michael Fox	Oscar Knox	Grand Total
2018	138437	141614	127145	135455	126344	176838	155111	157207	1158151
2019	105244	134764	114049	120302	105444	99493	96679	94465	870440
Grand Total	**243681**	**276378**	**241194**	**255757**	**231788**	**276331**	**251790**	**251672**	**2028591**

- **Visualization**



4.Revenue in each region

- **Introduction**

  By performing this analysis, we will get the sum of Revenue in each Region.

- **Description**

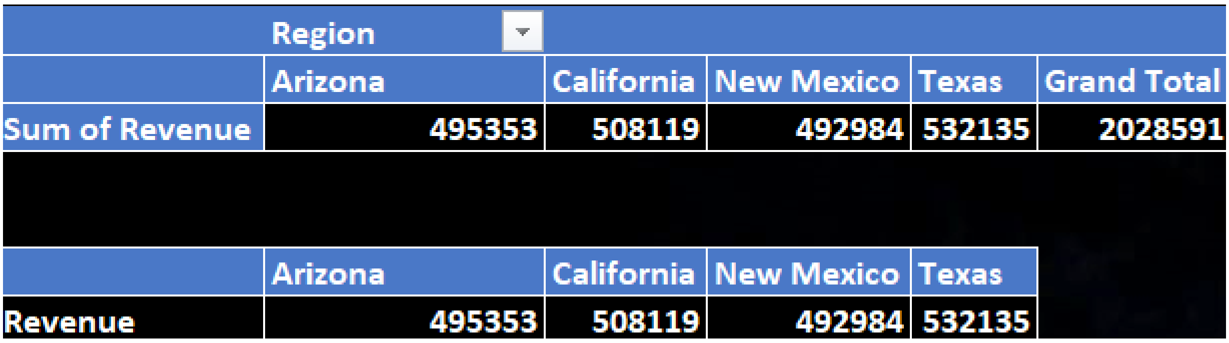
  The analysis based on Region, Revenue.

- **Specific requirements, functions and formulas**

  Pivot table is used for the analysis.

  Sum function is used in pivot table for the sum of the revenues in the pivot table

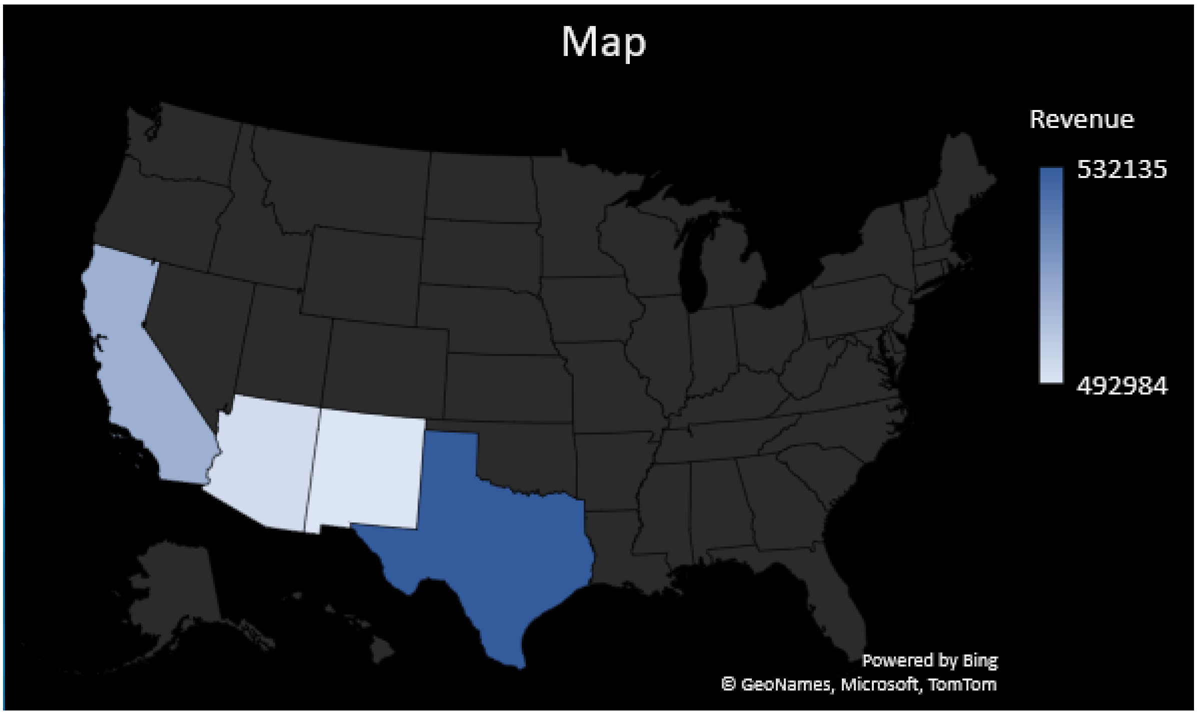
- **Analysis results**



	Region				
	Arizona	California	New Mexico	Texas	Grand Total
Sum of Revenue	495353	508119	492984	532135	2028591

	Arizona	California	New Mexico	Texas
Revenue	495353	508119	492984	532135

- **Visualization**



5.Sales trend in every month

- **Introduction**

  By performing this analysis, we will get the Trend of sales in each month in given years.

- **Description**

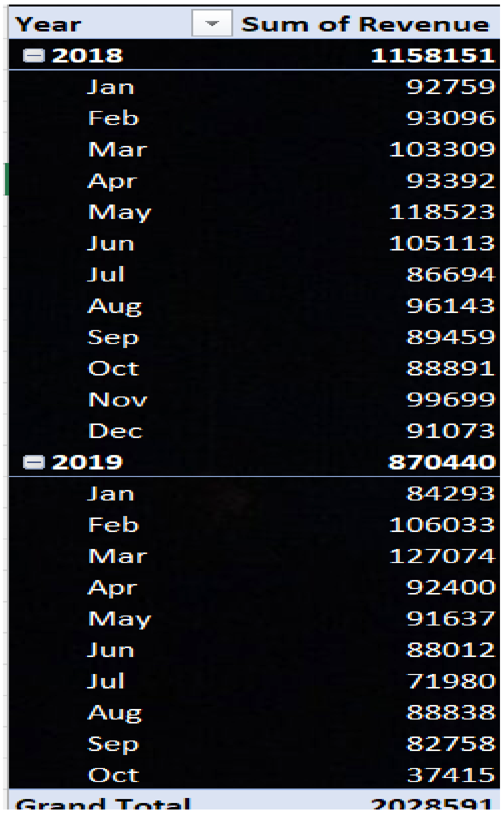
  The analysis based on Revenue, Date

- **Specific requirements, functions and formulas**

  Pivot table is used for the analysis.

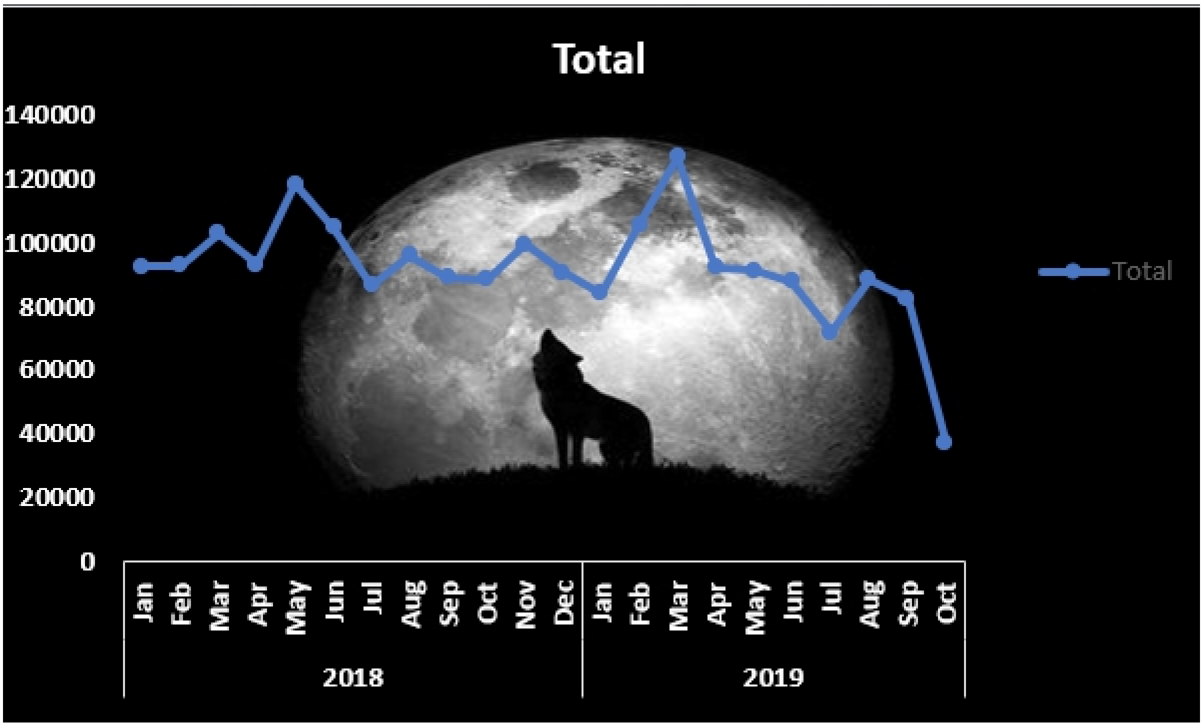
  Sum function is used in pivot table for the sum of the revenues in the pivot table

- **Analysis results**



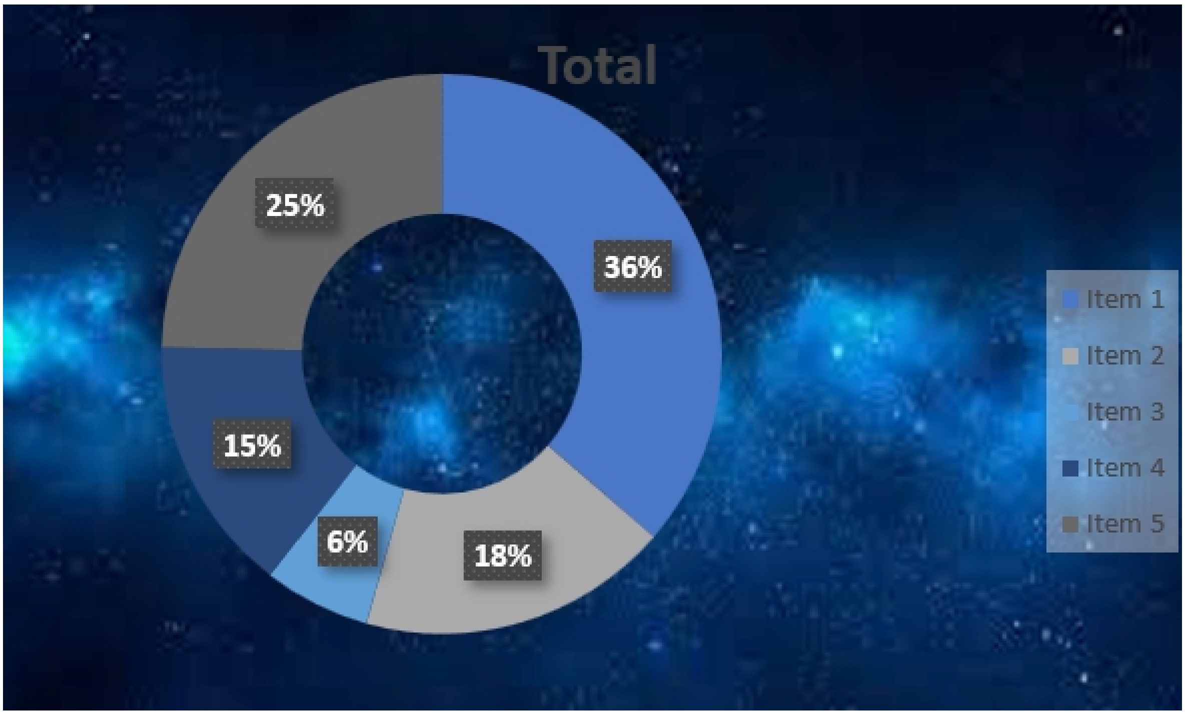
Year	Sum of Revenue
⊟2018	1158151
Jan	92759
Feb	93096
Mar	103309
Apr	93392
May	118523
Jun	105113
Jul	86694
Aug	96143
Sep	89459
Oct	88891
Nov	99699
Dec	91073
⊟2019	870440
Jan	84293
Feb	106033
Mar	127074
Apr	92400
May	91637
Jun	88012
Jul	71980
Aug	88838
Sep	82758
Oct	37415
Grand Total	2028591

- **Visualization**



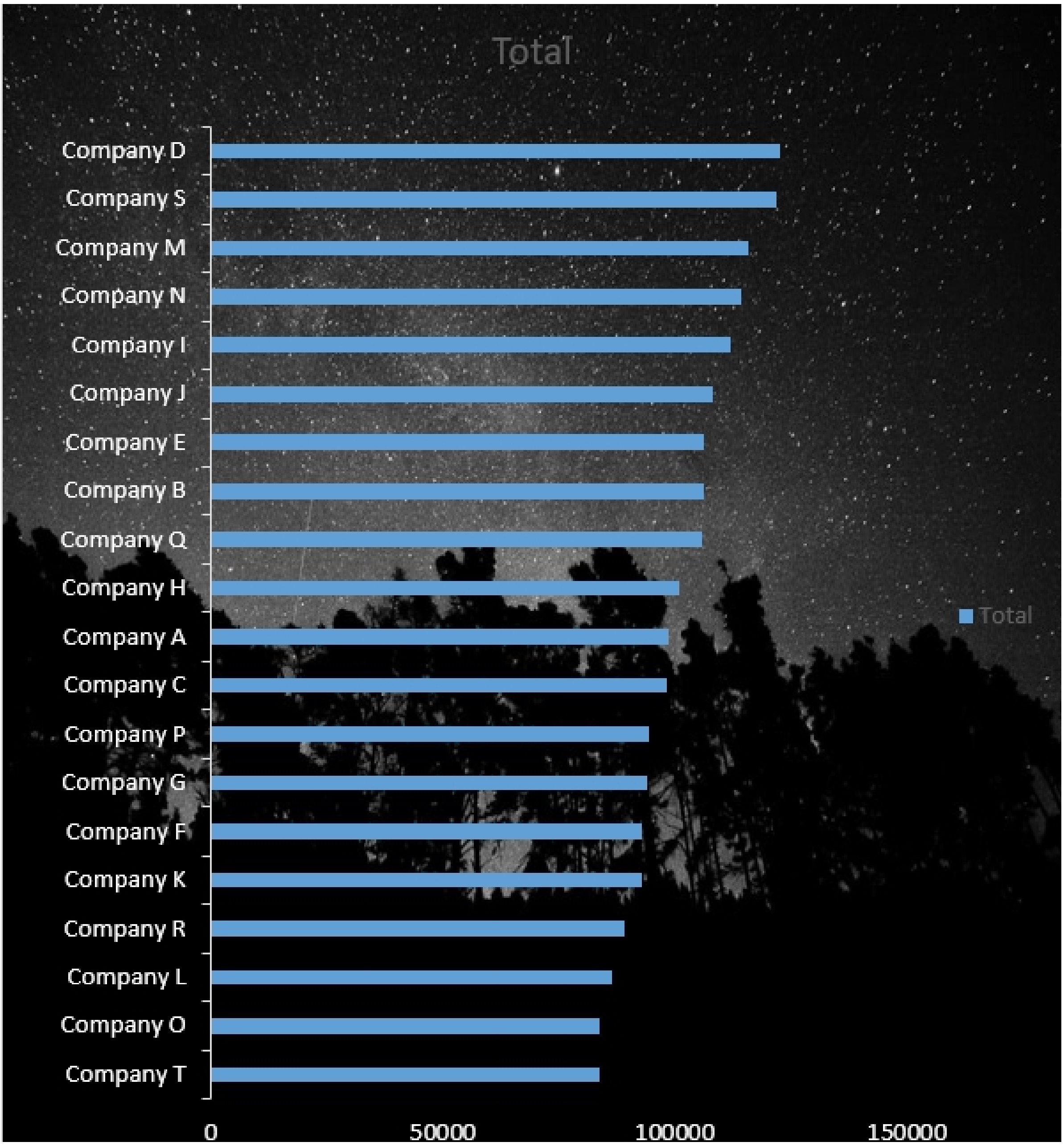
List of Analysis with results

- **Total Revenue through each item**
  - Item 1 have the highest revenue which is 36% it means customers are more interested to buy item 1.
  - Item 3 have the lowest revenue which is 6%.

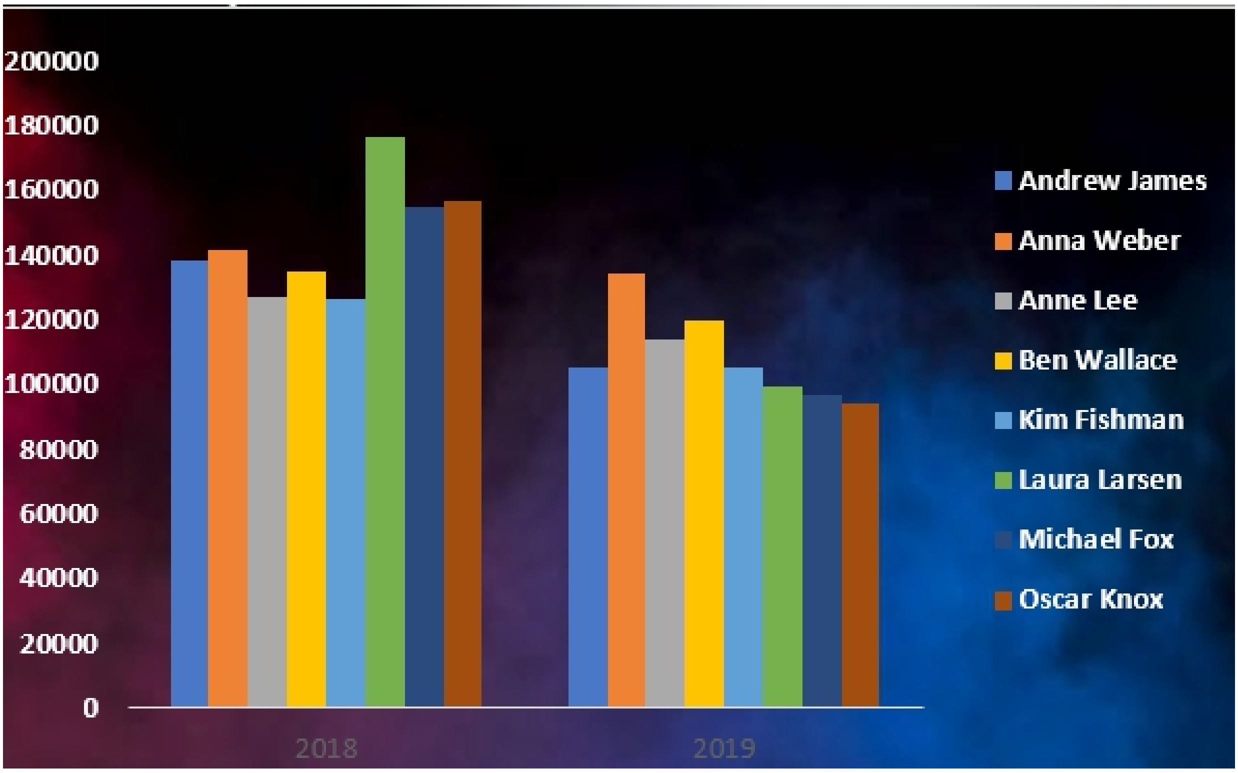


- **Revenue of each company**

  - Comapany D have the highest Revenue which means customers are more interested to buy the products which are manufactured in this company
  - Comapany T have the lowest Revenue among all.

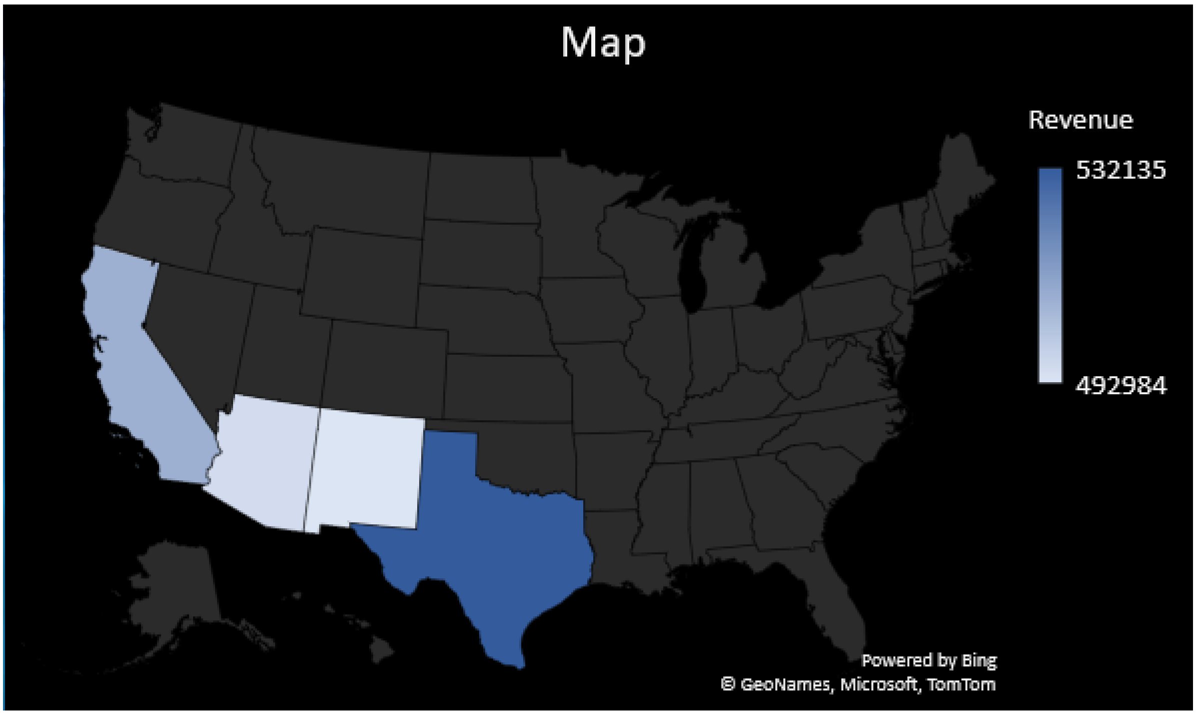


- **Sales by each employee in each year**
  - The year 2018 has more sales than in the tear 2019
  - The employee named Laura Larsen sold highest no of items in the year 2018 whereas the employee named kim fishman sold the least no of item.
  - The employee named Anna Weber sold highest no of items in the year 2019 whereas the employee named Oscar knox sold the least no of item.



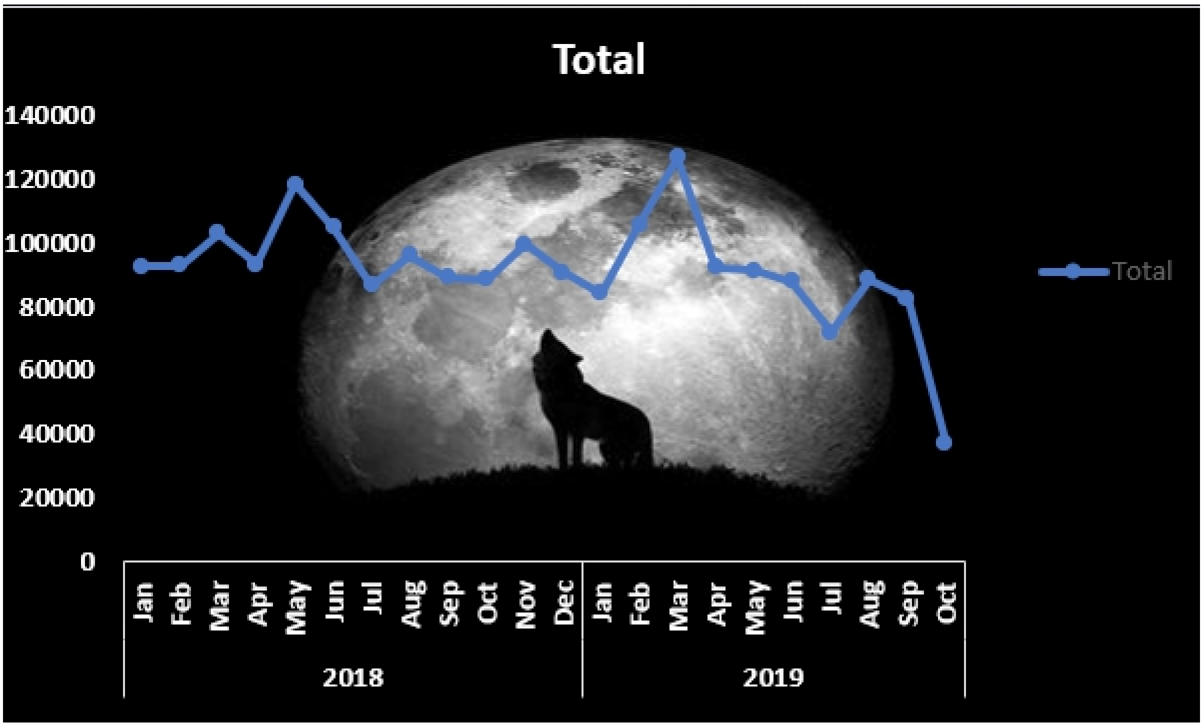
- **Revenue in each region**

  - Texas have the highest Revenue Compared to three of them. Which means sales persentage is highest in texas.
  - New mexico have the lowest Revenue.

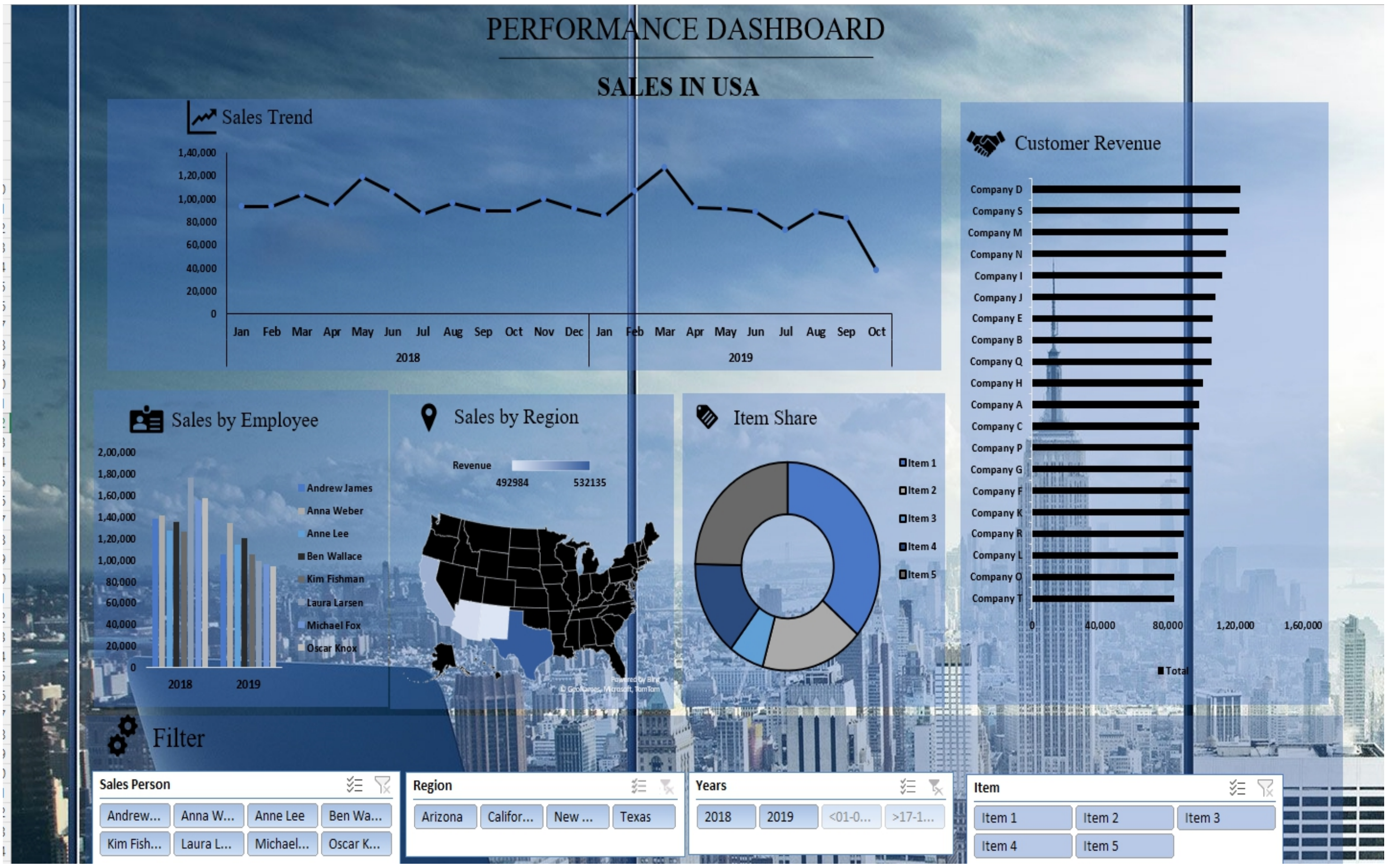


- **Sales trend in every month**

    - This analysis shows the sales trend in each month in two years
    - In the year 2018 the highest Revenue was in May and lowest Revenue was in October
    - In the year 2019 the highest Revenue was in March which is the greatest revenue in both years and lowest Revenue was in October



FINAL DASHBOARD



BIBLIOGRAPHY:

- **Dataset source:**

  https://www.kaggle.com/datasets

- **Dashboard Background Image:**

  https://www.pexels.com/search/hd%20background/

- **Information about Data Management:**

  What Is Data Management And Why It Is Vital | Blue-Pencil