**INT217: Introduction to Data Science**

****

**Project: Excel Dashboard on Internet Service Offers Dataset**

**Lovely Professional University, Phagwara**

**Faculty: Ms. Sameeksha Khare (27946)**

**Submitted by**

Ritu Priya Singh

Registration: 12001595 Roll No.: A01

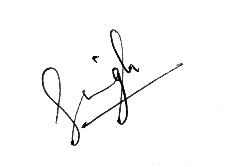
**Section: KM101**

**Project Duration: Semester 5 August – December 2022**

**DECLARATION**

I, Ritu Priya Singh, student of Lovely Professional University pursuing BTech CSE from School of Computer Science and Engineering, Punjab, hereby declare that all the information provided in this project report is based on my own analysis and is genuine.

Date: 9th Nov 2022 Signature



Registration No.: 12001595 Name of the Student

Ritu Priya Singh

**ACKNOWLEDGEMENT**

I would like to thank Ms. Sameeksha Khare Ma’am for her support and guidance in completing our project on the topic “Internet Service Offers Dataset”. It was a great learning experience for me in this semester.

Ritu Priya Singh

12001595

**INDEX**

* Introduction
* Objectives/ Scope of the Analysis
* Source of dataset
* ETL Process
* Analysis of Dataset
* List of Analysis with results
* Final Dashboard
* References

**Introduction**

This project (i.e., Dashboard) is developed to analyse the data of Internet Service Offers Dataset across dozens of US cities.

**Dashboard Objectives:**

1. To analyse the total Price of Internet Service & population with respect to Cities.
2. To analyse the technologies with respect to fastest speed price.
3. To analyse the count of upload speed and download speed with respect to packages.
4. To analyse the speed unit according to the population.
5. To analyse the average of latitude and longitude.
6. To analyse the average income with respect to cities.

**SOURCE OF DATASET**

**Link:** <https://www.kaggle.com/datasets/michaelbryantds/internet-speeds-and-prices>

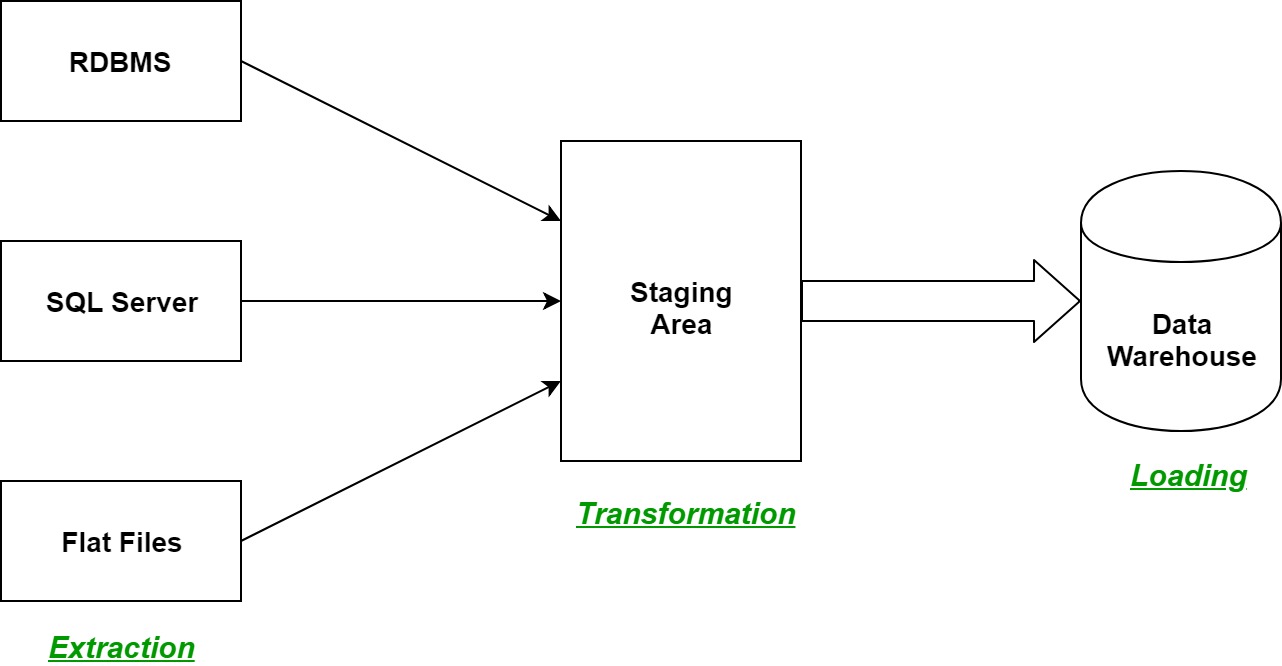
I choose my dataset from Kaggle.com as it is one of the best websites where different types of datasets are available for free and I have got one of the informative datasets for my project based on the Internet Service Offers.

**About Kaggle:**

Kaggle is an online community platform for data scientists and machine learning enthusiasts. Kaggle allows users to collaborate with other users, find and publish datasets, use GPU integrated notebooks, and compete with other data scientists to solve data science challenges. The aim of this online platform (founded in 2010 by Anthony Goldbloom and Jeremy Howard and acquired by Google in 2017) is to help professionals and learners reach their goals in their data science journey with the powerful tools and resources it provides. As of today (2021), there are over 8 million registered users on Kaggle.

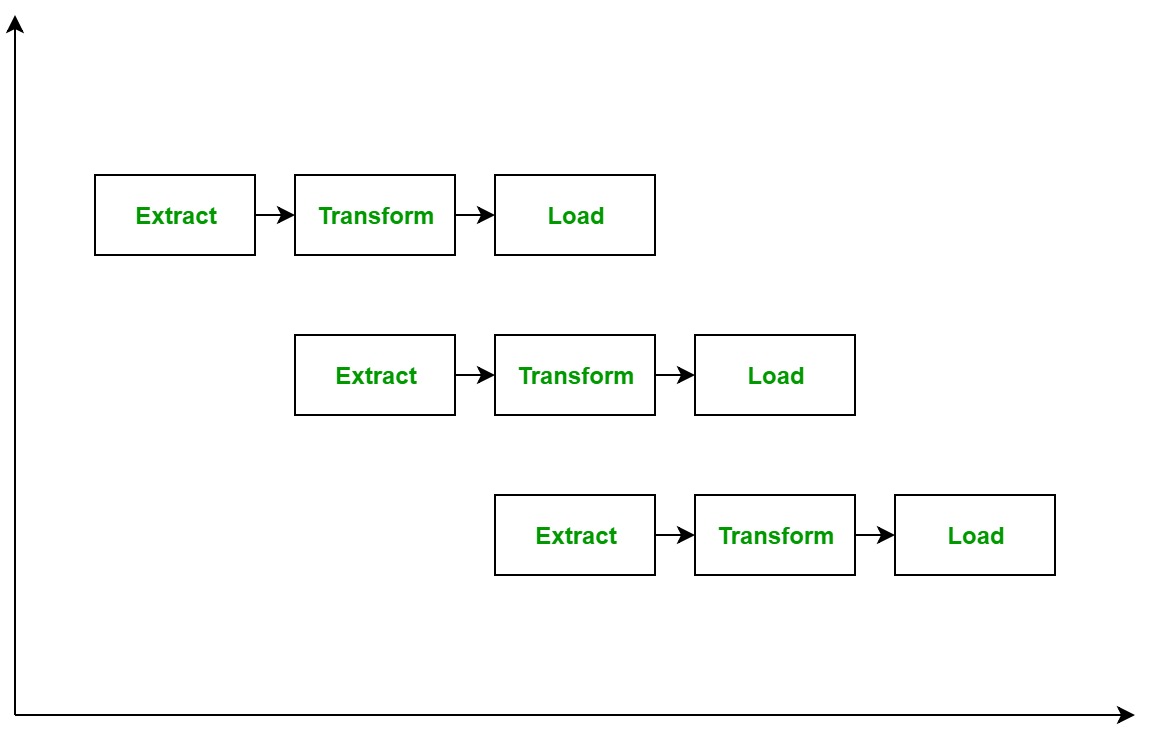
**ETL PROCESS**

ETL is a process in Data Warehousing and it stands for Extract, Transform and Load. It is a process in which an ETL tool extracts the data from various data source systems, transforms it in the staging area, and then finally, loads it into the Data Warehouse system.



* **Extraction:** 
  + The first step of the ETL process is extraction. In this step, data from various source systems is extracted which can be in various formats like relational databases, No SQL, XML, and flat files into the staging area. It is important to extract the data from various source systems and store it into the staging area first and not directly into the data warehouse because the extracted data is in various formats and can be corrupted also. Hence loading it directly into the data warehouse may damage it and rollback will be much more difficult. Therefore, this is one of the most important steps of ETL process.
* **Transformation:**
  + The second step of the ETL process is transformation. In this step, a set of rules or functions are applied on the extracted data to convert it into a single standard format.
  + Filtering – loading only certain attributes into the data warehouse.
  + Cleaning – filling up the NULL values with some default values, mapping U.S.A, United States, and America into USA, etc.
  + Joining – joining multiple attributes into one.
  + Splitting – splitting a single attribute into multiple attributes.
  + Sorting – sorting tuples on the basis of some attribute (generally key-attribute).
* **Loading:**
  + The third and final step of the ETL process is loading. In this step, the transformed data is finally loaded into the data warehouse. Sometimes the data is updated by loading into the data warehouse very frequently and sometimes it is done after longer but regular intervals. The rate and period of loading solely depends on the requirements and varies from system to system.

ETL process can also use the pipelining concept i.e., as soon as some data is extracted, it can transform and during that period some new data can be extracted. And while the transformed data is being loaded into the data warehouse, the already extracted data can be transformed. The block diagram of the pipelining of ETL process is shown below:

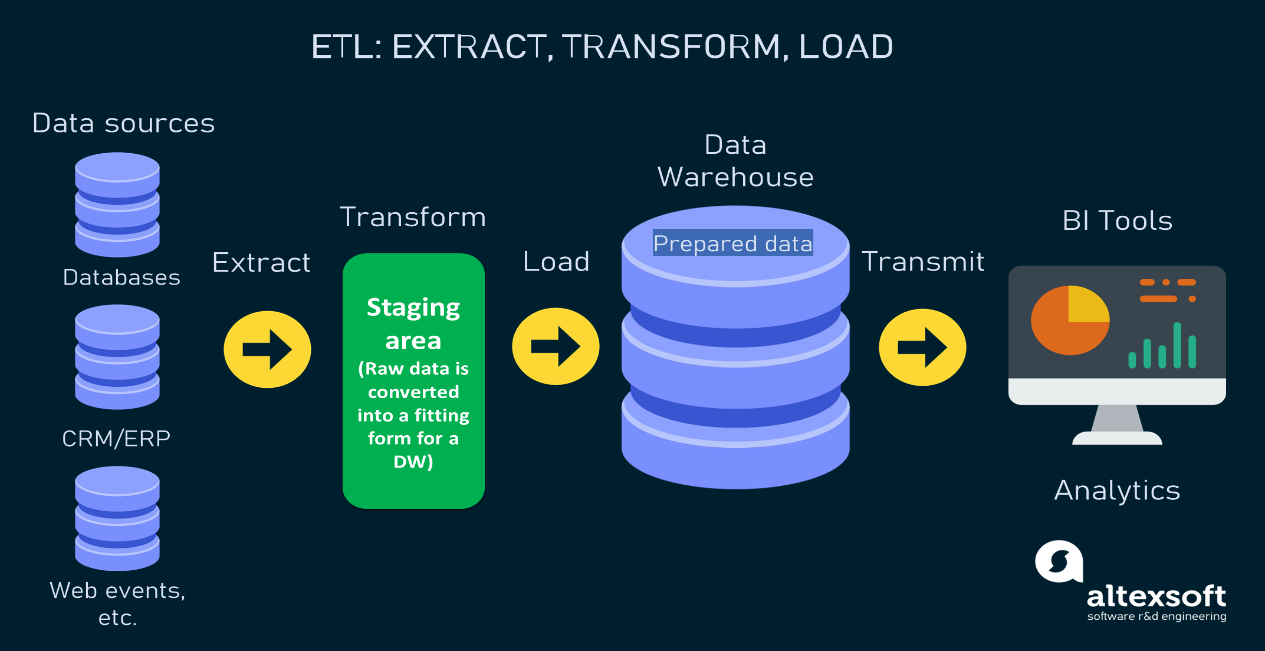


ETL Tools: Most commonly used ETL tools are Hevo, Sybase, Oracle Warehouse builder, CloverETL, and MarkLogic.

Data Warehouses: Most commonly used Data Warehouses are Snowflake, Redshift, BigQuery, and Firebolt.

**The benefits and challenges of ETL:**

ETL solutions improve quality by performing data cleansing prior to loading the data to a different repository. A time-consuming batch operation, ETL is recommended more often for creating smaller target data repositories that require less frequent updating, while other data integration methods—including ELT (extract, load, transform), change data capture (CDC), and data virtualization—are used to integrate increasingly larger volumes of data that changes or real-time data streams.



**Analysis of Dataset**

***Dataset contains:***

* Number of columns: 27
* Number of rows: 432304

***Dataset after cleaning:***

* Number of columns: 19
* Number of rows: 432304

***Columns contain:***

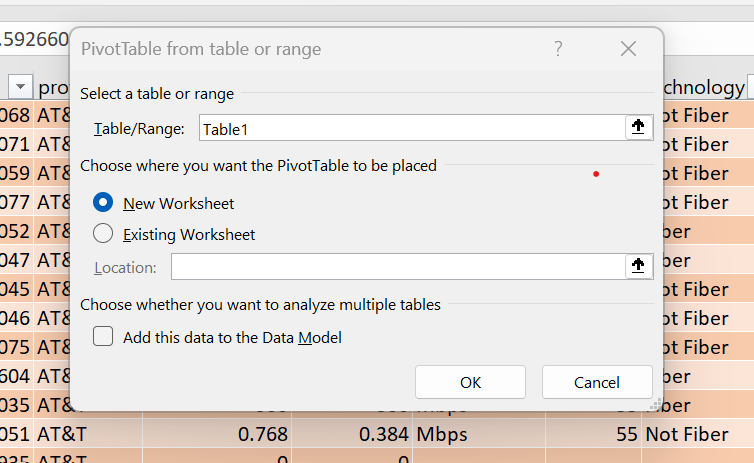
1. Major City
2. State
3. Latitude
4. Longitude
5. Provider: The internet service provider
6. Speed down: Cheapest advertised download speed for the address
7. Speed up: Cheapest advertised upload speed for the address
8. Speed unit: The unit of speed is in Mbps
9. Price: The cost in USD of the cheapest advertised internet plan
10. Technology: The kind of technology (fiber or non-fiber) used to serve the cheapest internet plan
11. Package: The name of the cheapest internet plan
12. Fastest speed down: The download speed of the fastest package
13. Fastest speed price: The upload speed of the fastest internet package
14. Redlining grade: The redlining grade, merged from Mapping Inequality based on the latitude and longitude of the address
15. Ppl per square mile: People per square mile is used to determine population density
16. N Providers: The number of other wired competitors in the addresses
17. Income dollars below median: City median household income – median household income
18. Internet perc broadband: The percentage of the population that is already subscribed to broadband
19. Median household income

***Specific Requirement:***

* MS Excel
* Pivot Table
* Slicer
* Power Point

***Steps to Creating the Pivot Table:***

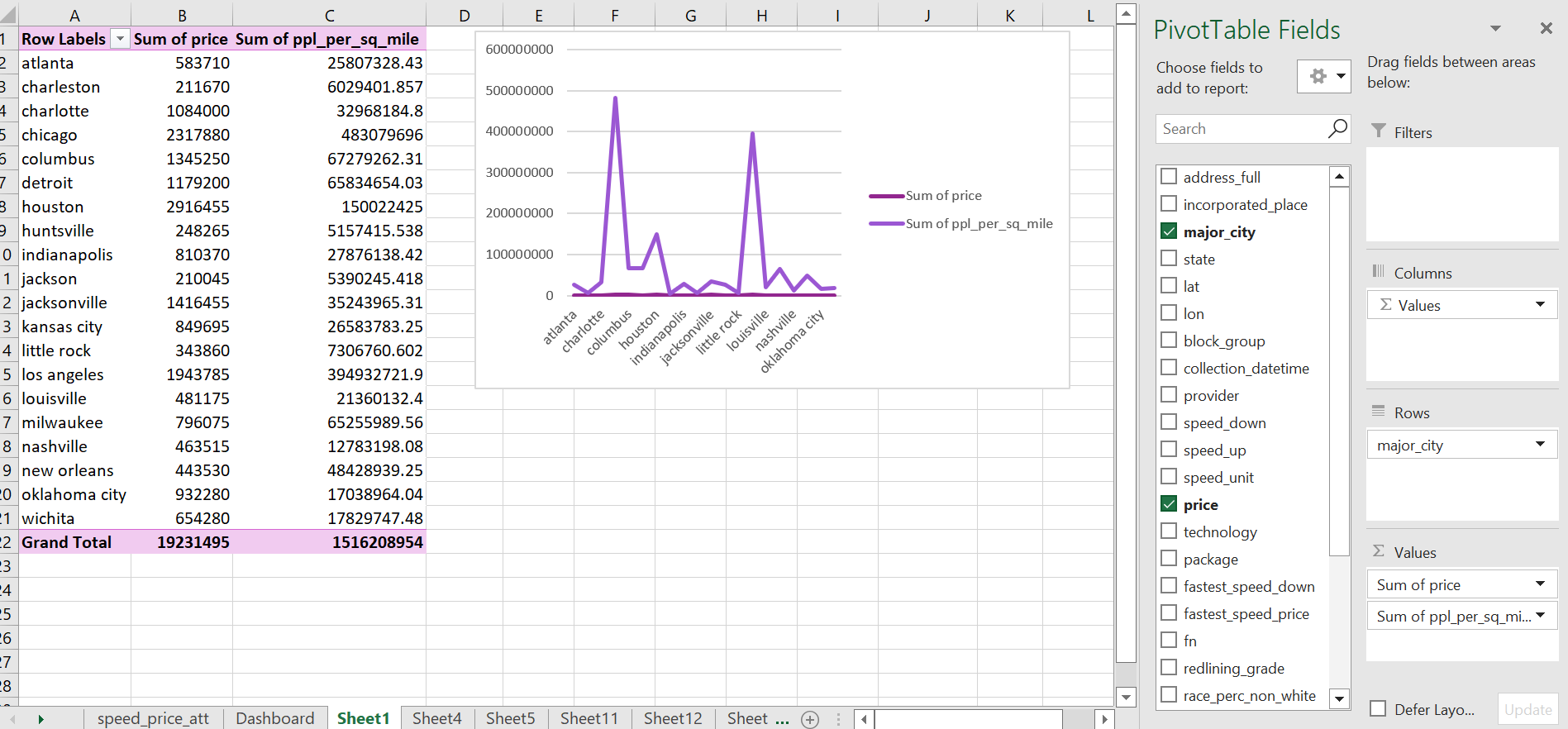
* Select a cell in the dataset.
* Go to Insert> Pivot Table
* 

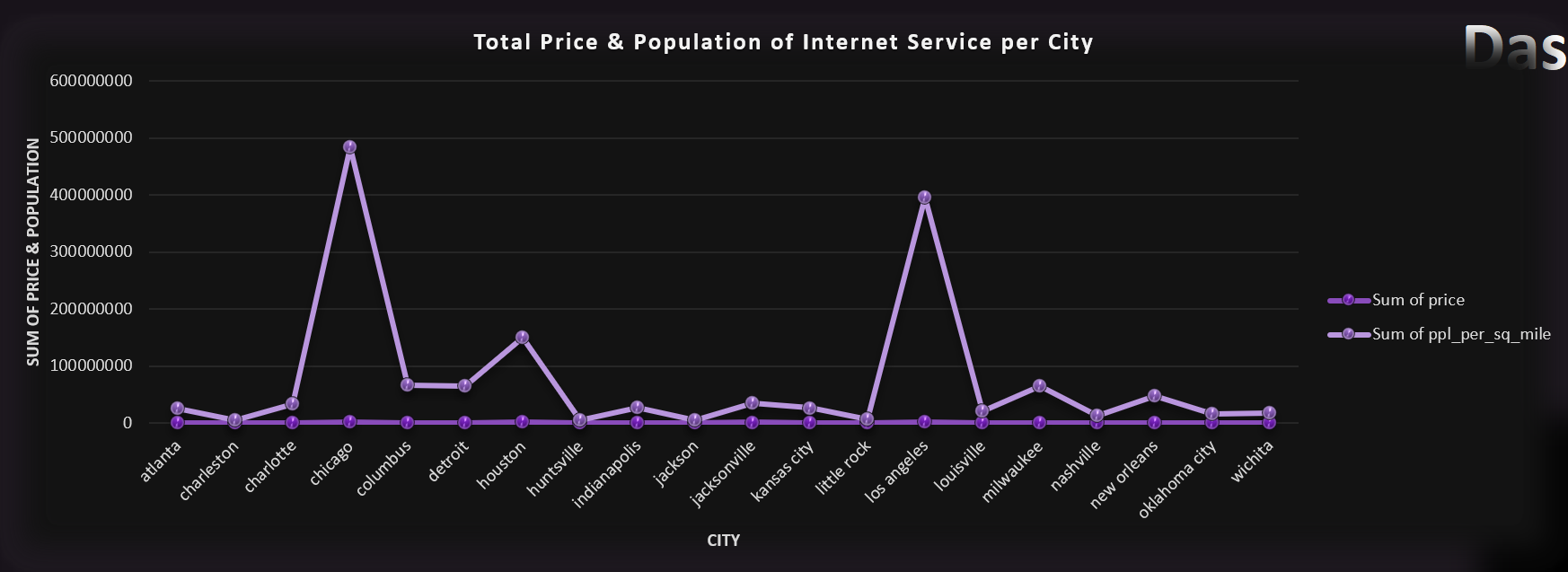


* Select New Worksheet.
* Provide table name or leave it as default.
* Click on OK.

**To analyse the total Price of Internet Service & population with respect to Cities:**

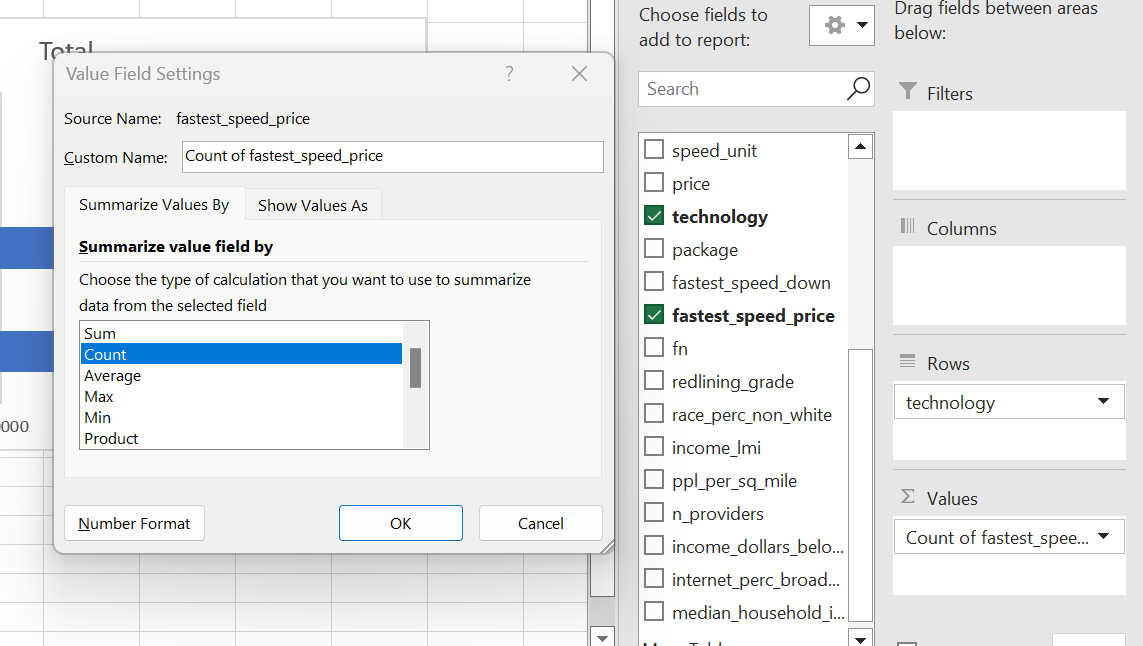
* This table contains the sum of the price of cheapest internet plan which is in US Dollar.
* Process of creating Line Chart using this Pivot Table.
  + Create Pivot Table> Select major\_city as rows> Select price & ppl\_per\_sq\_mile in the Values section.
  + Go to Insert> Select Line chart in the Charts section.



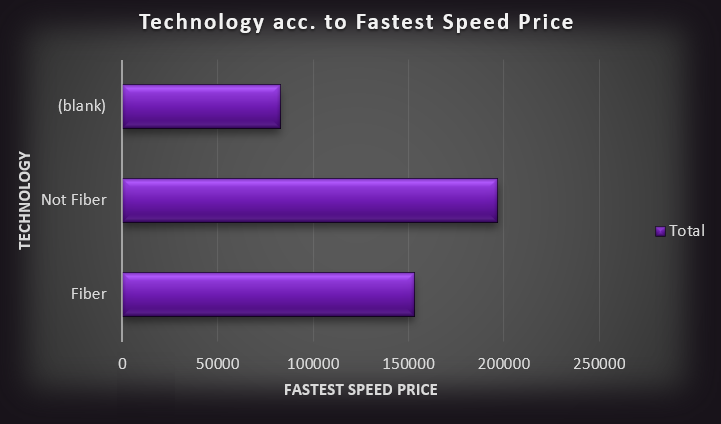
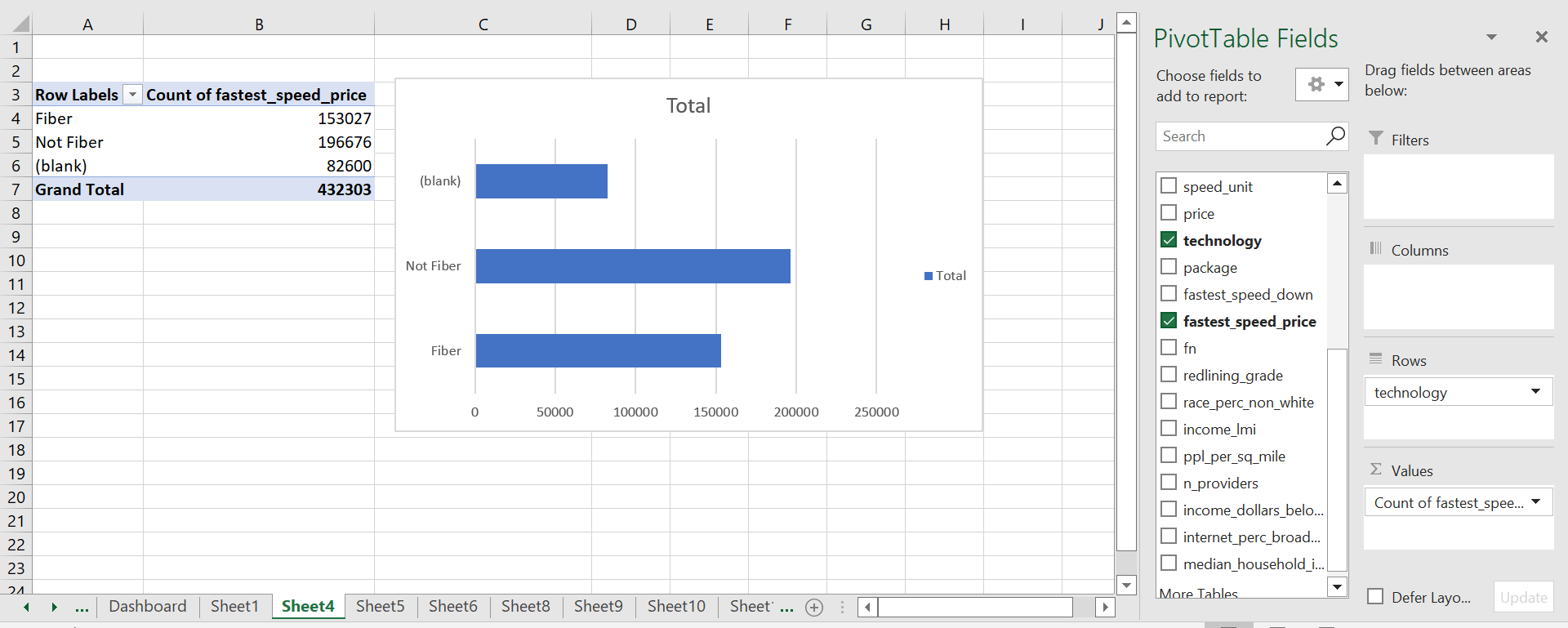


**To analyse the technologies with respect to fastest speed price:**

* Create Pivot Table> Select Technology as rows> fastest\_speed\_price as values.

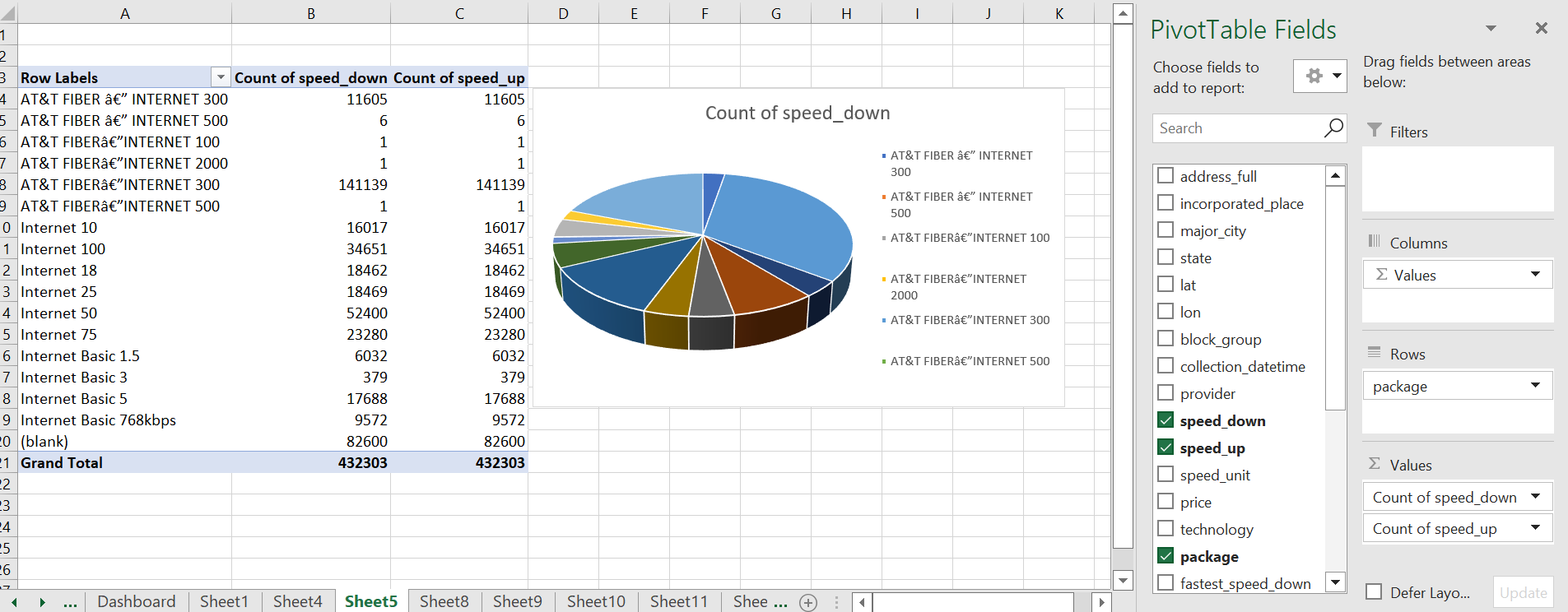


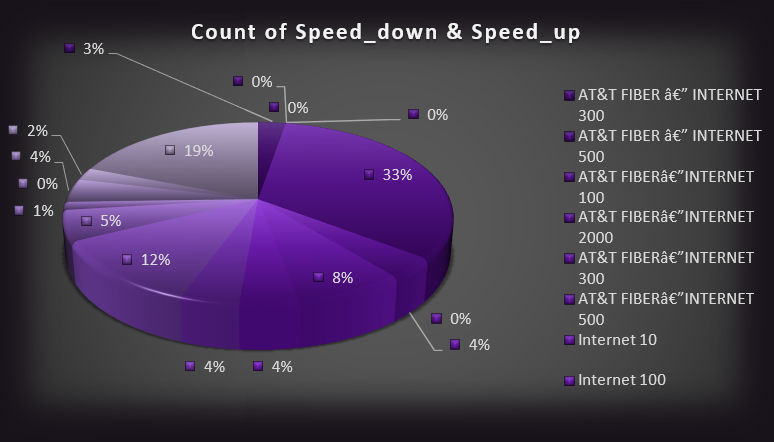
* Click on Values section> Value field settings> select Count option > OK.
* Go to Insert> Select 2D bar chart type> OK.



**To analyse the count of upload speed and download speed with respect to packages:**

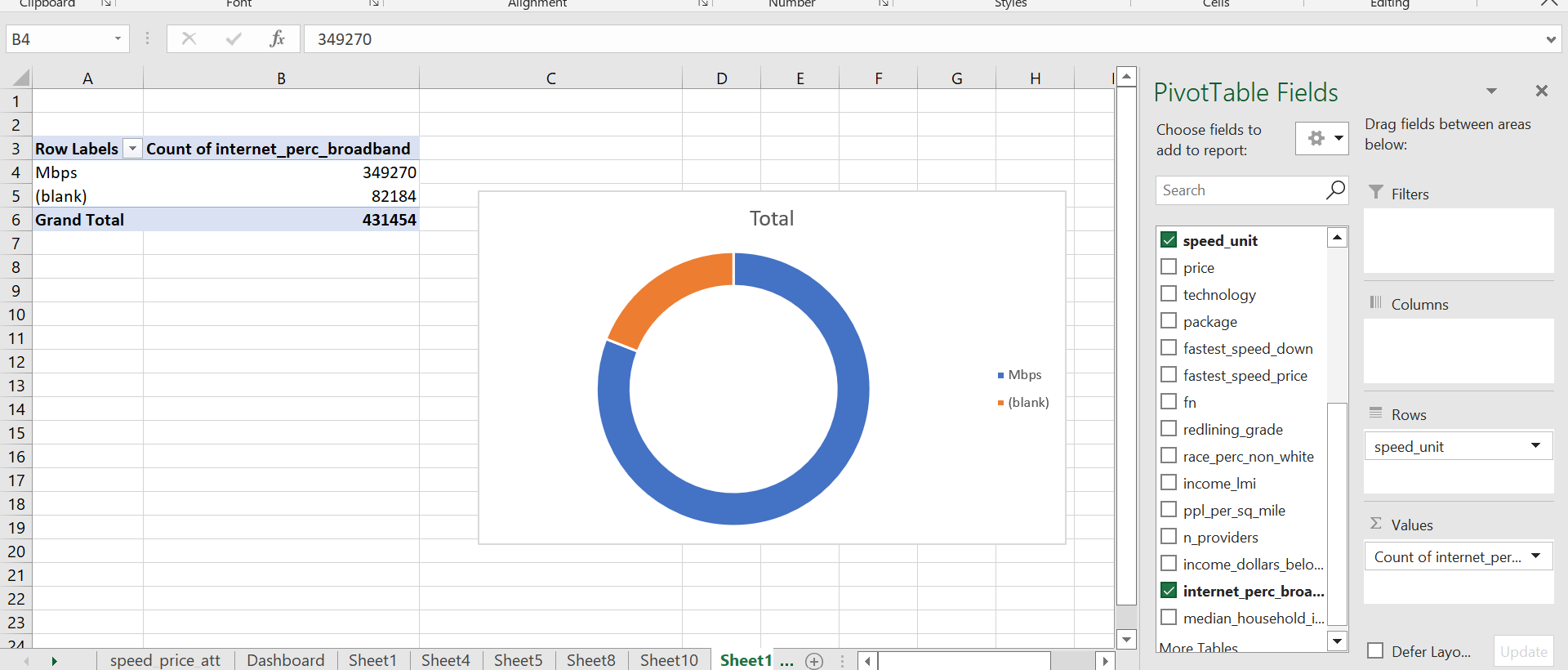
* Create Pivot table> select speed\_up & speed\_down in the values section> select package in the rows.
* Insert> Select 3D pie chart from the chart section> click OK.

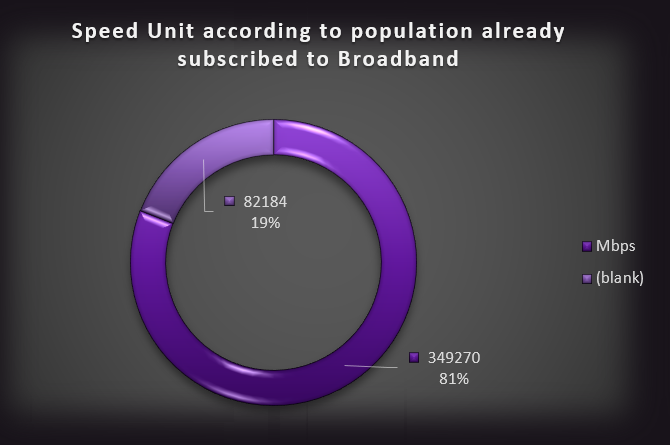




**To analyse the speed unit according to the population:**

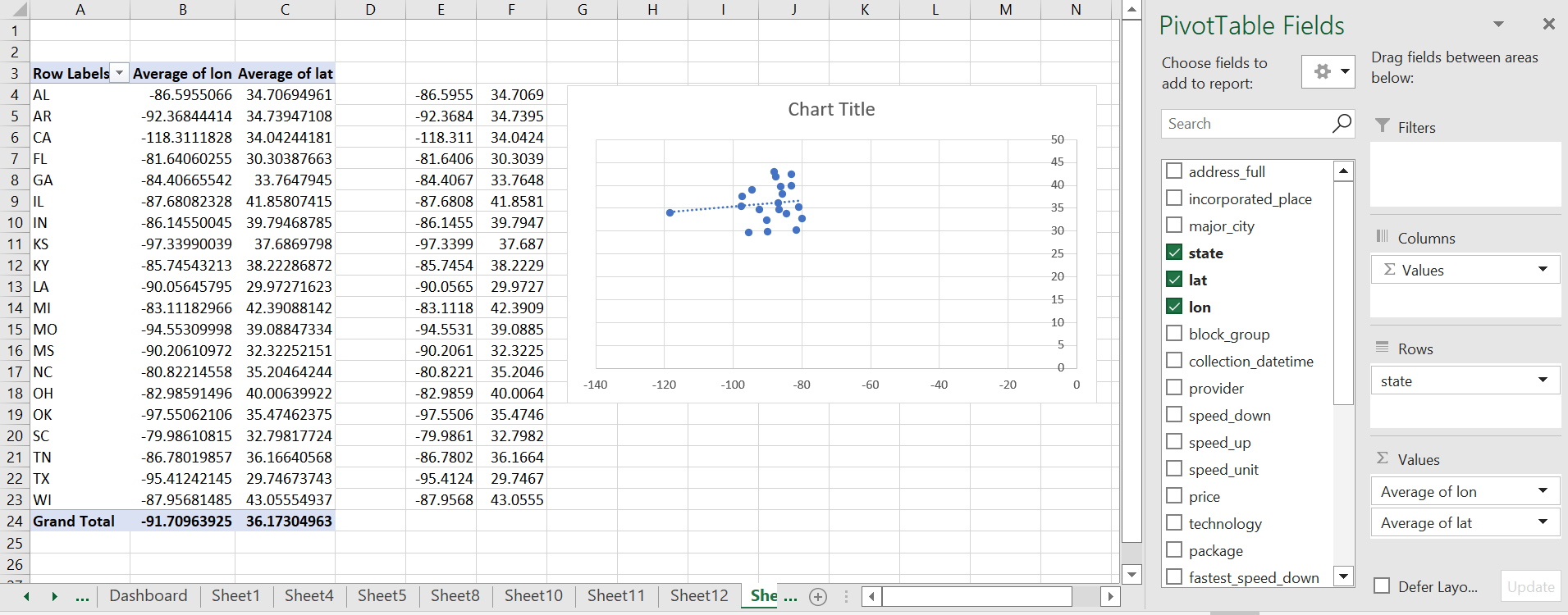
* Create Pivot Table> select speed\_unit in the rows area> select internet\_perc\_broadband in the values area.
* Edit the values field settings> select count> Click OK.
* Select Donut chart from the chart section.
* Click OK.

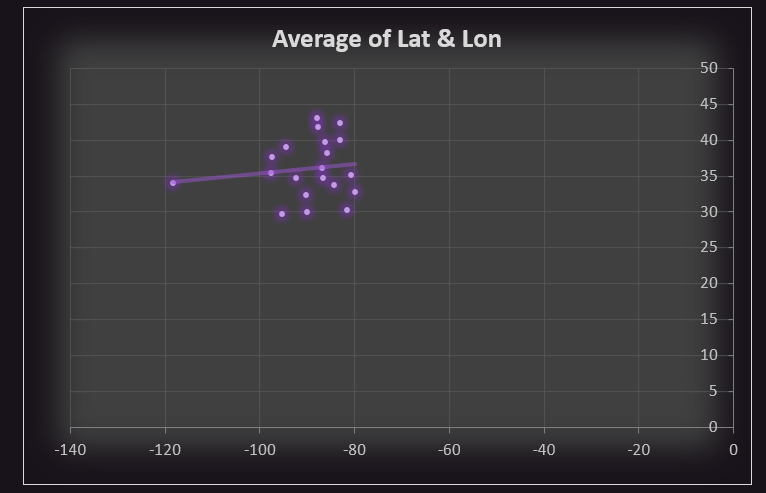




**To analyse the average of latitude and longitude:**

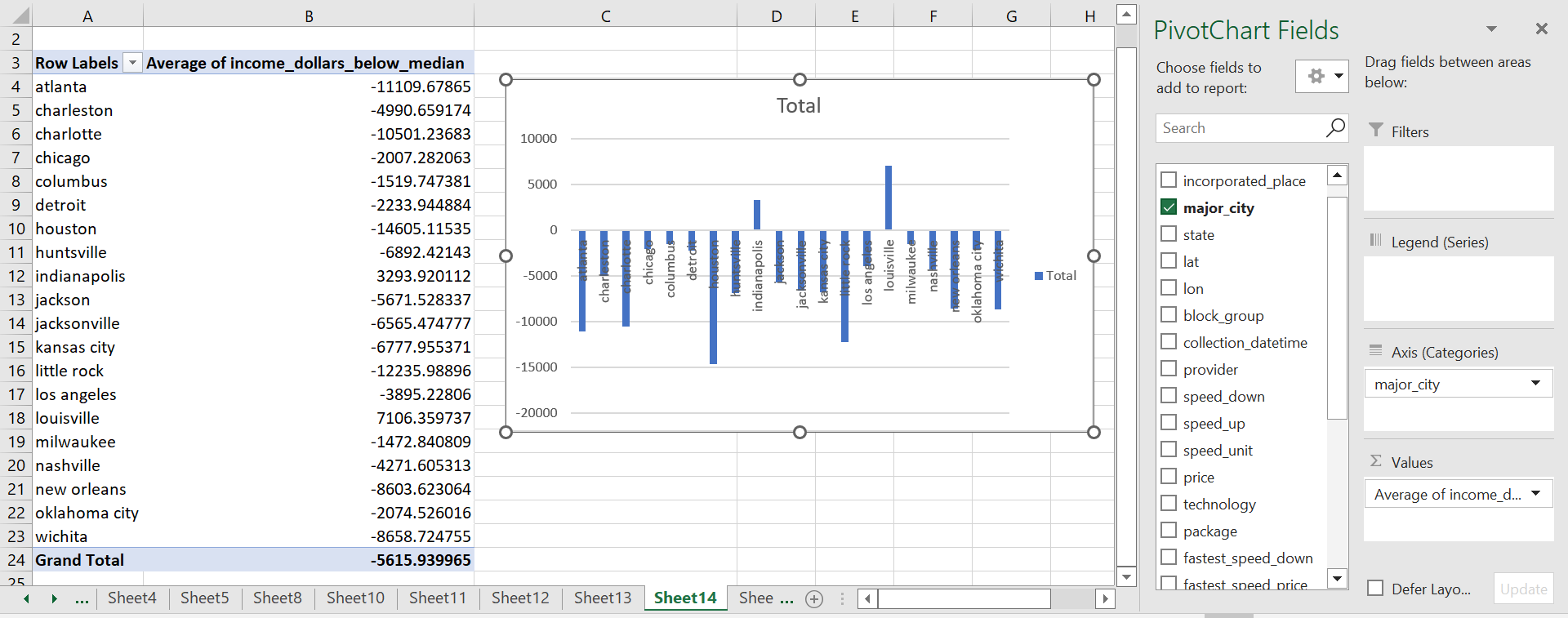
* Create Pivot table> select state in the rows> select the values of lat and lon in the columns area.
* In the values field> select the average of latitude and average of longitude.
* Click OK.
* Copy the values of latitude and longitude from the pivot table and paste it in another cell.
* Select one of the cells> go to insert> chart option> click on Scatter chart> click OK.

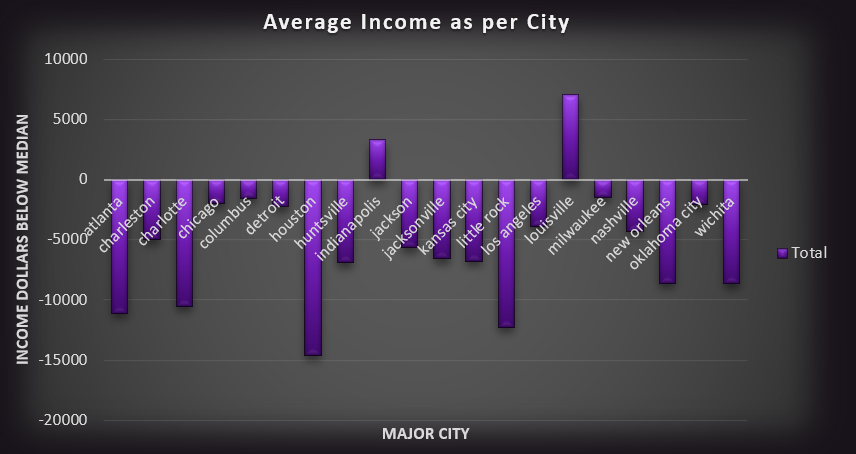




**To analyse the average income with respect to cities**

* Create Pivot table> select major city in the rows section> select income\_dollars\_below\_median in the values section.
* Edit Value field settings> select average option> click OK.
* Select bar chart from the charts option.
* Click OK.





**Slicers:**

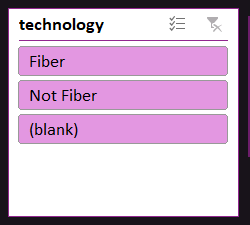
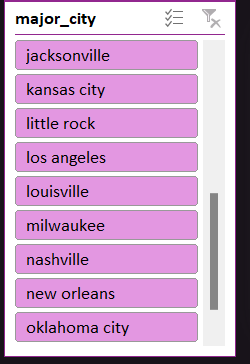
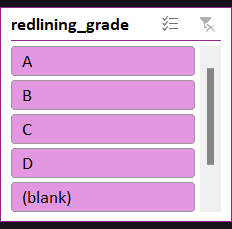
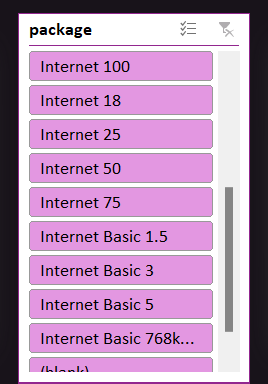
Slicers provide buttons that you can click to filter tables, or PivotTables. In addition to quick filtering, slicers also indicate the current filtering state, which makes it easy to understand what exactly is currently displayed.

**How to use Slicer:**

1. Click anywhere in the table or PivotTable.
2. On the Home tab, go to Insert> Slicer.

Insert Slicer

1. In the Insert Slicers dialog box, select the check boxes for the fields you want to display, then select OK.
2. A slicer will be created for every field that you selected. Clicking any of the slicer buttons will automatically apply that filter to the linked table or PivotTable.
3. You can adjust your slicer preferences in the Slicer tab on the ribbon.
4. If you want to connect a slicer to more than one PivotTable, go to Slicer> Report Connections> check the PivotTables to include, then select OK.

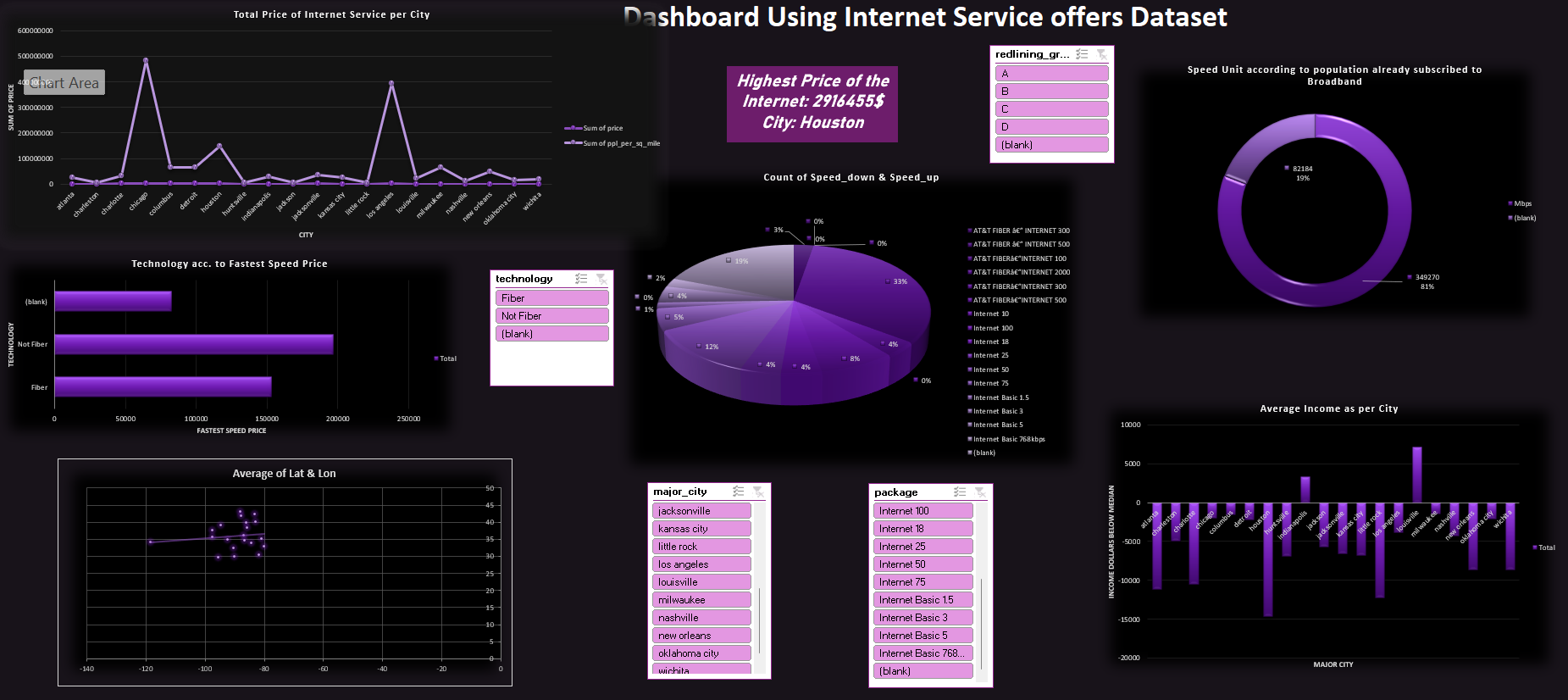
****

**Dashboard:**

A dashboard is a visual display of all of your data. While it can be used in all kinds of different ways, its primary intention is to provide information at-a-glance, such as KPIs.

A dashboard usually sits on its own page and receives information from a linked database. In many cases it’s configurable, allowing you the ability to choose which data you want to see and whether you want to include charts or graphs to visualize the numbers.

**Overview of the Dashboard**



**Summary:**

* The excel dashboard is used to display overviews of the large datasets.
* Excel dashboard use the elements like pivot tables, charts, graphs and gauges to show the overviews.
* The dashboard eases the decision-making process by showing the vital parts of the data in the same place.

**Key Features:**

* Total price of the Internet.
* Technologies which are available: fiber & non-fiber.
* Total Population density per square mile.
* Upload speed and download speed.
* Speed unit which measures in Mbps only.
* Fastest download speed of the fastest package available.
* Income in dollars.

**Special Features:**

* We can analyse the dataset according to the cities in US, redlining grade, technology, and package of the internet service provider.
* Using the slicer in Cities can show the data of a particular city.
* Package in slicer can show the total price of that package in the cities.
* Technology in slicer can show the count of median household income.

**References**

* https://www.kaggle.com/
* https://support.microsoft.com/en-us/office/use-slicers-to-filter-data-249f966b-a9d5-4b0f-b31a-12651785d29d#:~:text=Slicers%20provide%20buttons%20that%20you,what%20exactly%20is%20currently%20displayed.

***THANK YOU***