

# Day2 Content

- Pandas Data Frame and Series
- Data Analysis using Pandas
- Linear Regression using Ordinary Least Square
- Azure ML Studio Basics
- Linear Regression Model using Azure ML Studio

# Data Storage Formats in Pandas

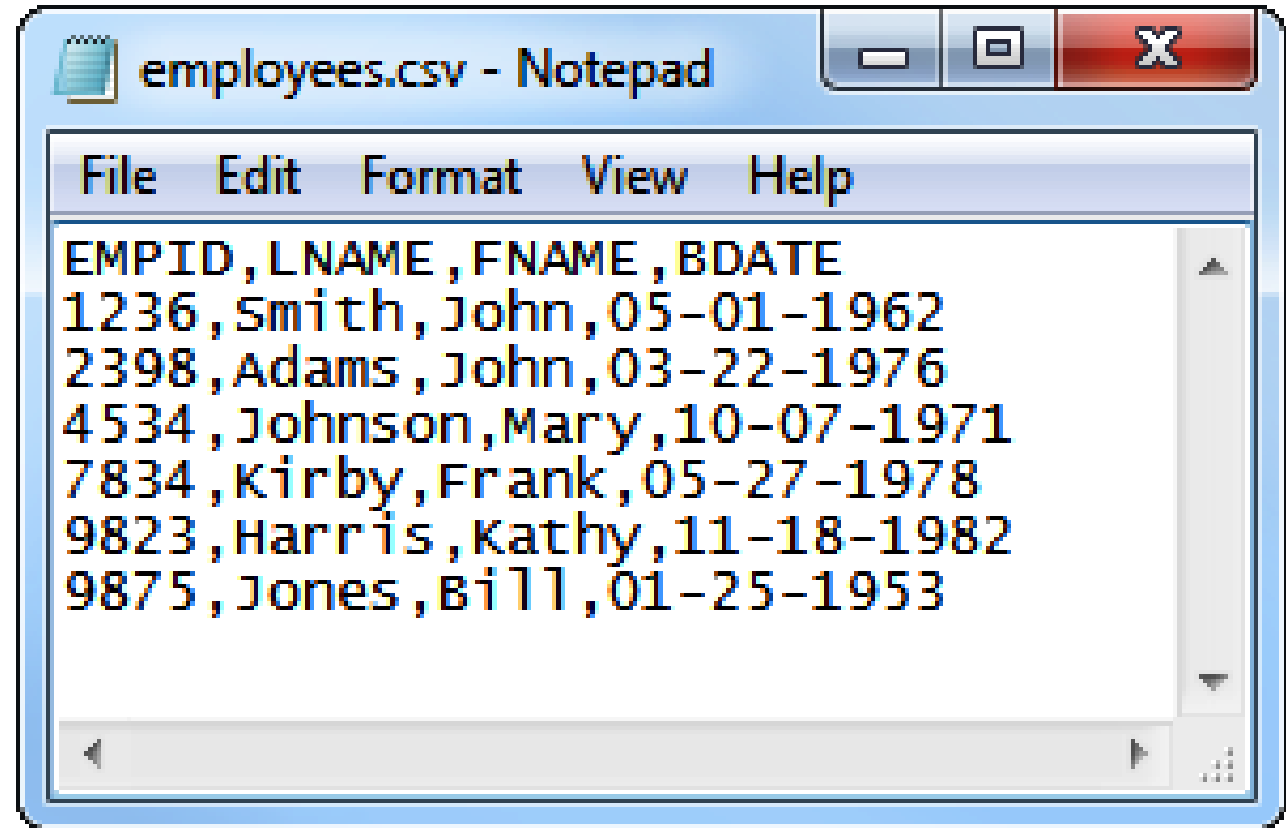
The different data storage formats available to be manipulated by Pandas library are text, binary and SQL.

| Format Type | Data Description     | Reader         | Writer       |
|-------------|----------------------|----------------|--------------|
| text        | CSV                  | read_csv       | to_csv       |
| text        | JSON                 | read_json      | to_json      |
| text        | HTML                 | read_html      | to_html      |
| text        | Local clipboard      | read_clipboard | to_clipboard |
| binary      | MS Excel             | read_excel     | to_excel     |
| binary      | HDF5 Format          | read_hdf       | to_hdf       |
| binary      | Feather Format       | read_feather   | to_feather   |
| binary      | Parquet Format       | read_parquet   | to_parquet   |
| binary      | Msgpack              | read_msgpack   | to_msgpack   |
| binary      | Stata                | read_stata     | to_stata     |
| binary      | SAS                  | read_sas       |              |
| binary      | Python Pickle Format | read_pickle    | to_pickle    |
| SQL         | SQL                  | read_sql       | to_sql       |
| SQL         | Google Big Query     | read_gbq       | to_gbq       |

Reference: <https://www.programmersought.com/article/4943826582>

# What is CSV file?

A CSV is a comma-separated values file, which allows data to be saved in a tabular format.



Reference: <https://www.goanywhere.com/managed-file-transfer/more/tutorials/how-to-import-csv-file-into-database>

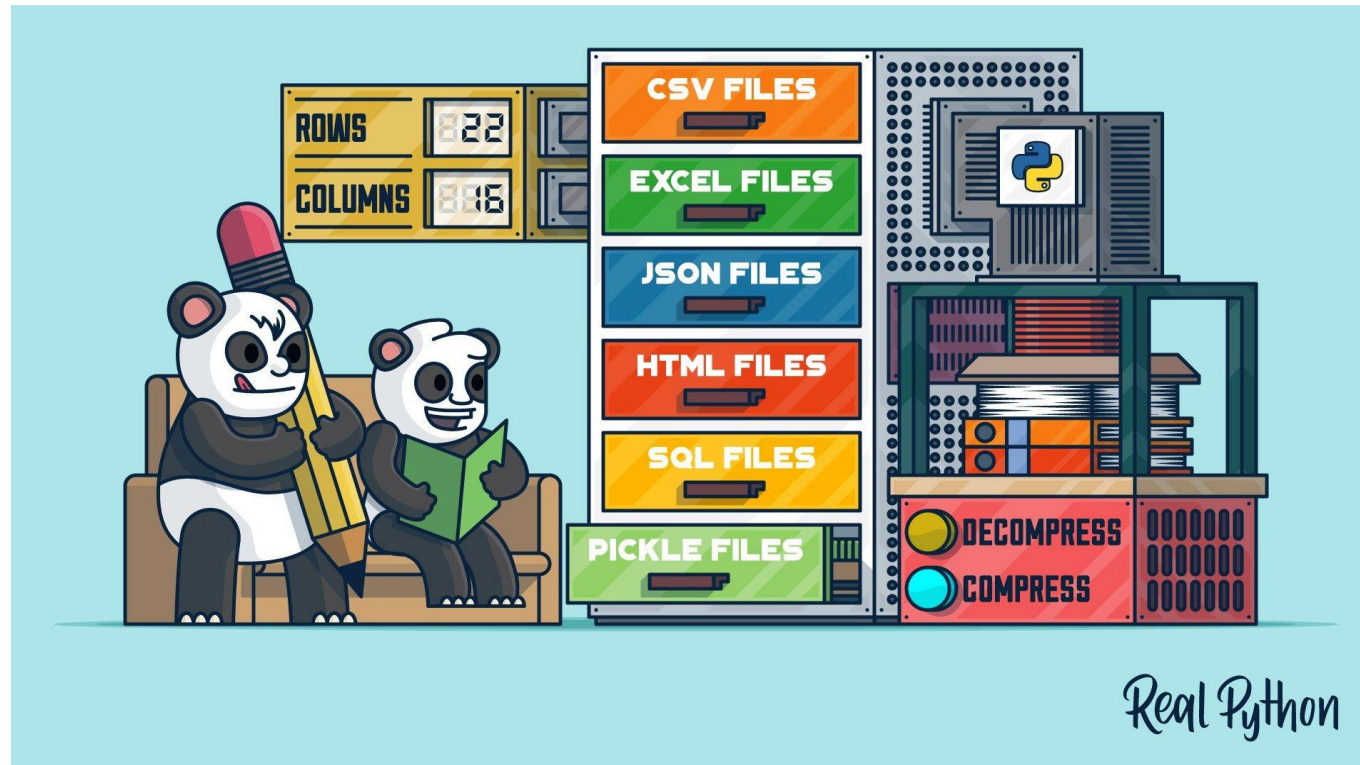
# What is a JSON file?

A JSON file is a file that stores simple data structures and objects in JavaScript Object Notation (JSON) format, which is a standard data interchange format.

```
{
  "Product": {
    "0": "Desktop Computer",
    "1": "Tablet",
    "2": "iPhone",
    "3": "Laptop"
  },
  "Price": {
    "0": 700,
    "1": 250,
    "2": 800,
    "3": 1200
  }
}
```

Reference: <https://datatofish.com/export-pandas-dataframe-json>

# Reading data from files



Reference: <https://realpython.com/pandas-read-write-files>

# Load CSV files to Python Pandas

The basic process of loading data from a CSV file into a Pandas DataFrame is achieved using the “read\_csv” function in Pandas.

```
In [18]: pd.read_csv("../pokemon.csv", header=[6,3,5,7], squeeze = True)
```

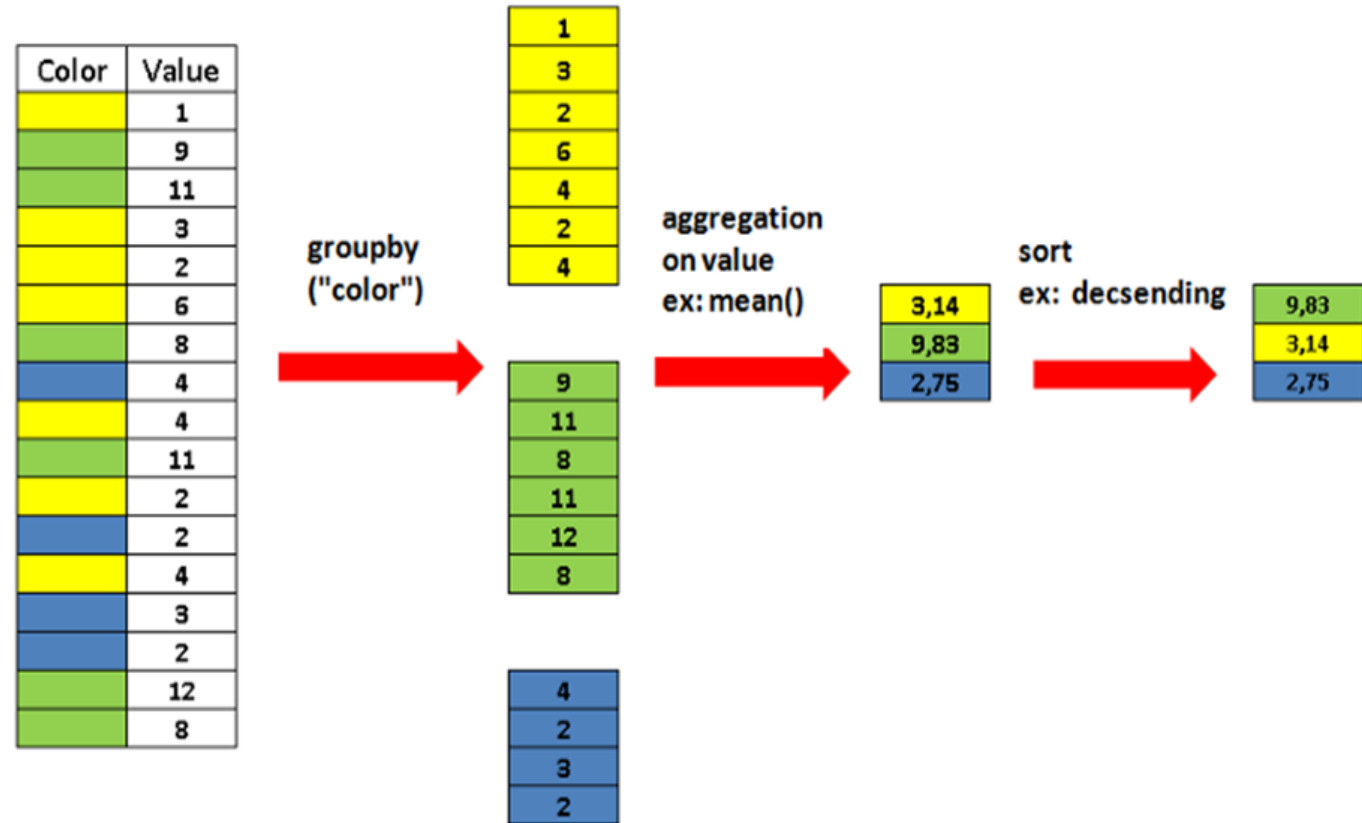
```
Out[18]:
```

|   |            |       |
|---|------------|-------|
|   | Charizard  | Fire  |
|   | Venusaur   | Grass |
|   | Charmeleon | Fire  |
|   | Squirtle   | Water |
| 0 | Wartortle  | Water |
| 1 | Blastoise  | Water |
| 2 | Caterpie   | Bug   |
| 3 | Metapod    | Bug   |

Reference: [https://www.geeksforgeeks.org/python-read-csv-using-pandas-read\\_csv](https://www.geeksforgeeks.org/python-read-csv-using-pandas-read_csv)

# Groupby Methods

Pandas `dataframe.groupby()` function is used to split the data into groups based on some criteria. pandas objects can be split on any of their axes.



Reference: <https://towardsdatascience.com/pandas-groupby-explained-453692519d0>



# Groupby output format – Series or DataFrame?

As a rule of thumb, if you calculate more than one column of results, your result will be a DataFrame. For a single column of results, the agg function, by default, will produce a Series.

```
In [35]: data.groupby('month', as_index=False).agg({"duration": "sum"})
```

```
Out[35]:
```

|   | month   | duration  |
|---|---------|-----------|
| 0 | 2014-11 | 26639.441 |
| 1 | 2014-12 | 14641.870 |
| 2 | 2015-01 | 18223.299 |
| 3 | 2015-02 | 15522.299 |
| 4 | 2015-03 | 22750.441 |

Reference: <https://www.shanelynn.ie/summarising-aggregation-and-grouping-data-in-python-pandas>

# Pivot Tables

It's a table of statistics that helps summarize the data of a larger table by “pivoting” that data.

Pivot

df

|   | foo | bar | baz | zoo |
|---|-----|-----|-----|-----|
| 0 | one | A   | 1   | x   |
| 1 | one | B   | 2   | y   |
| 2 | one | C   | 3   | z   |
| 3 | two | A   | 4   | q   |
| 4 | two | B   | 5   | w   |
| 5 | two | C   | 6   | t   |

→

```
df.pivot(index='foo',
          columns='bar',
          values='baz')
```

| bar | A | B | C |
|-----|---|---|---|
| foo |   |   |   |
| one | 1 | 2 | 3 |
| two | 4 | 5 | 6 |

[https://pandas.pydata.org/pandas-docs/version/0.25.3/user\\_guide/reshaping.html](https://pandas.pydata.org/pandas-docs/version/0.25.3/user_guide/reshaping.html)

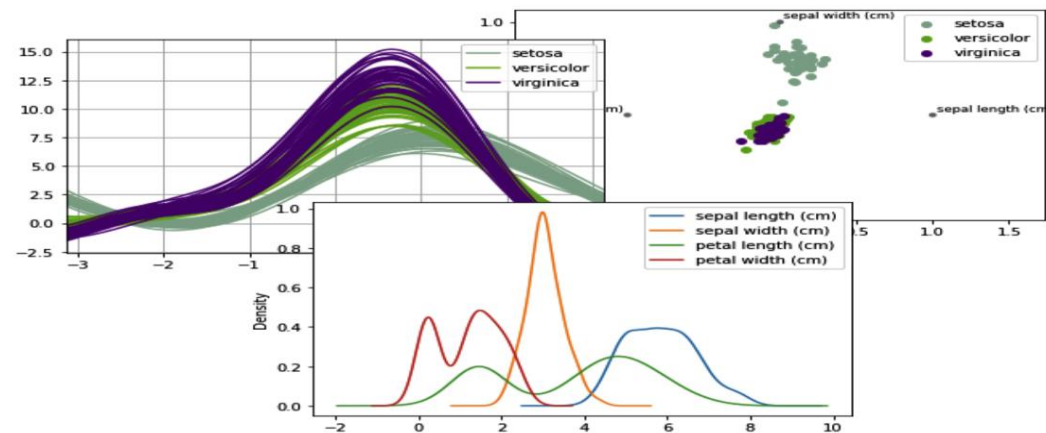
# How to Build a Pivot Table in Python

In Pandas, we can construct a pivot table using the following syntax:

```
pandas.pivot_table(data, values=None, index=None, columns=None,  
aggfunc='mean', fill_value=None, margins=False, dropna=True,  
margins_name='All', observed=False)
```

# Pandas Plotting

Plotting in pandas utilises the matplotlib API so in order to create visualisations, you will need to also import this library alongside pandas.



<https://towardsdatascience.com/the-best-pandas-plotting-features-c9789e04a5a0>

# Plot a Scatter Diagram using Pandas

Scatter plots are used to depict a relationship between two variables.

Step 1: Prepare the data

| Unemployment_Rate | Stock_Index_Price |
|-------------------|-------------------|
| 6.1               | 1500              |
| 5.8               | 1520              |
| 5.7               | 1525              |
| 5.7               | 1523              |
| 5.8               | 1515              |
| 5.6               | 1540              |
| 5.5               | 1545              |
| 5.3               | 1560              |
| 5.2               | 1555              |
| 5.2               | 1565              |

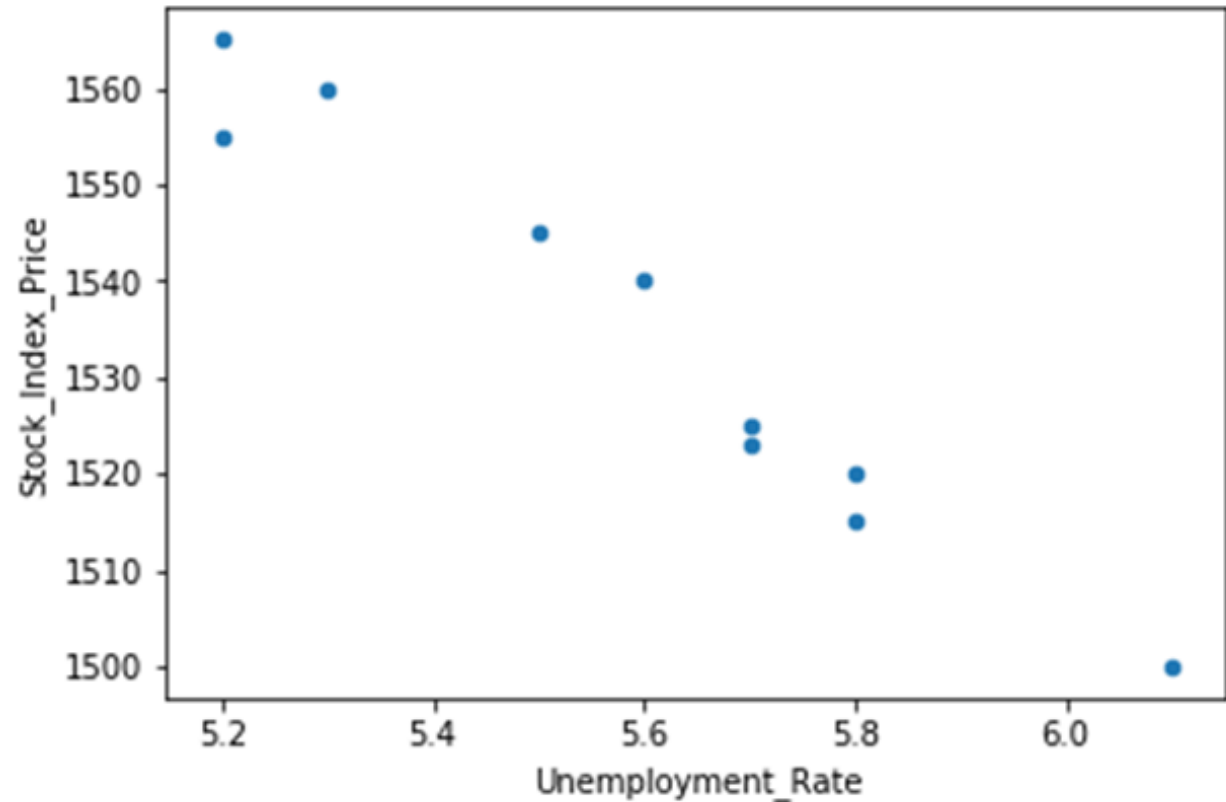
# Plot a Scatter Diagram using Pandas

Step 2: Create the DataFrame

|   | Unemployment_Rate | Stock_Index_Price |
|---|-------------------|-------------------|
| 0 | 6.1               | 1500              |
| 1 | 5.8               | 1520              |
| 2 | 5.7               | 1525              |
| 3 | 5.7               | 1523              |
| 4 | 5.8               | 1515              |
| 5 | 5.6               | 1540              |
| 6 | 5.5               | 1545              |
| 7 | 5.3               | 1560              |
| 8 | 5.2               | 1555              |
| 9 | 5.2               | 1565              |

# Plot a Scatter Diagram using Pandas

Step 3: Plot the DataFrame using Pandas



# Plot a Line Chart using Pandas

Line charts are often used to display trends overtime.

Step 1: Prepare the data

| Year | Unemployment_Rate |
|------|-------------------|
| 1920 | 9.8               |
| 1930 | 12                |
| 1940 | 8                 |
| 1950 | 7.2               |
| 1960 | 6.9               |
| 1970 | 7                 |
| 1980 | 6.5               |
| 1990 | 6.2               |
| 2000 | 5.5               |
| 2010 | 6.3               |



# Plot a Line Chart using Pandas

Step 2: Create the DataFrame

|   | Year | Unemployment_Rate |
|---|------|-------------------|
| 0 | 1920 | 9.8               |
| 1 | 1930 | 12.0              |
| 2 | 1940 | 8.0               |
| 3 | 1950 | 7.2               |
| 4 | 1960 | 6.9               |
| 5 | 1970 | 7.0               |
| 6 | 1980 | 6.5               |
| 7 | 1990 | 6.2               |
| 8 | 2000 | 5.5               |
| 9 | 2010 | 6.3               |

# Plot a Line Chart using Pandas

Step 3: Plot the DataFrame using Pandas



# Plot a Bar Chart using Pandas

Bar charts are used to display categorical data.

Step 1: Prepare the data

| Country | GDP_Per_Capita |
|---------|----------------|
| USA     | 45000          |
| Canada  | 42000          |
| Germany | 52000          |
| UK      | 49000          |
| France  | 47000          |

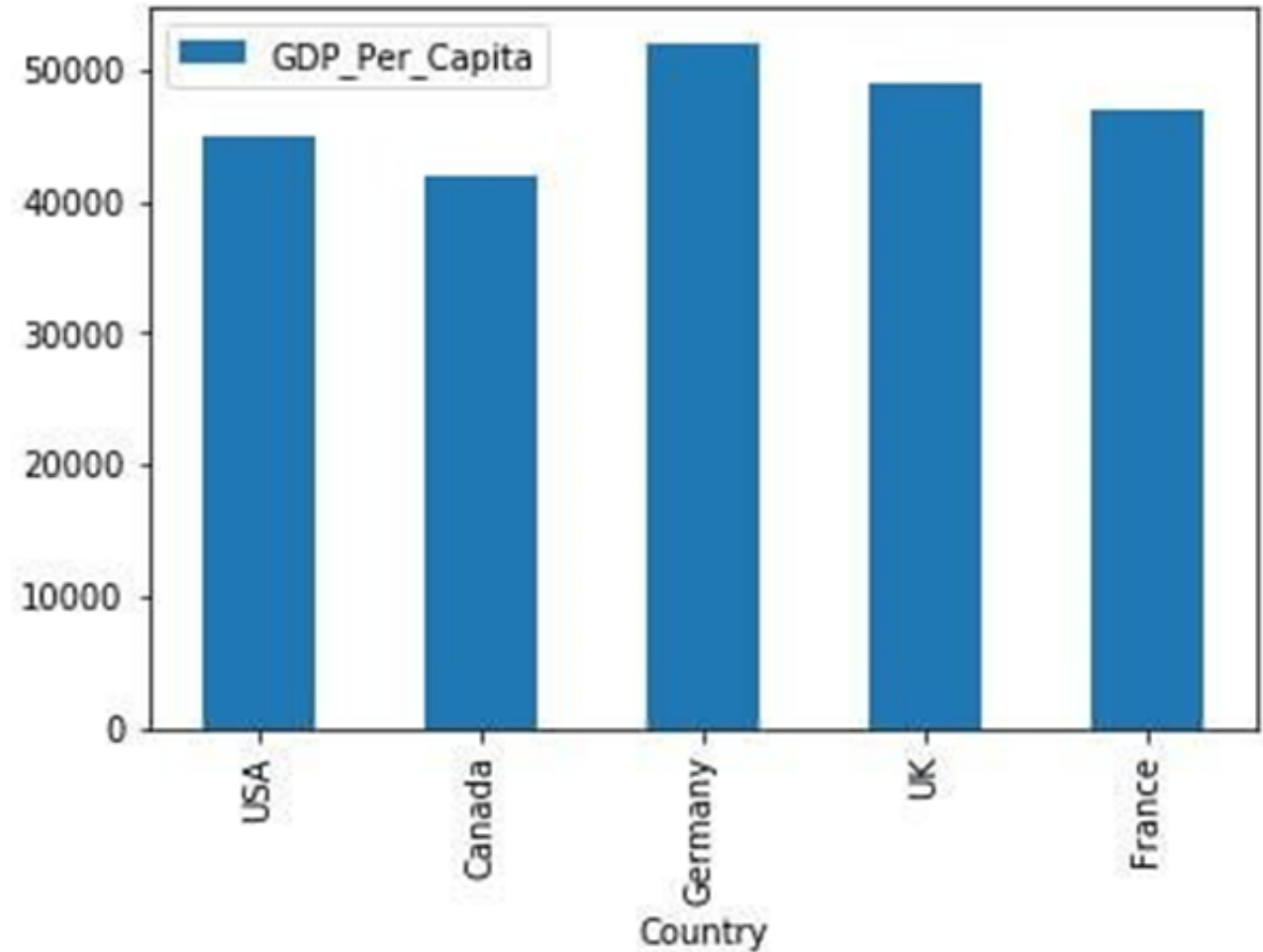
# Plot a Bar Chart using Pandas

Step 2: Create the DataFrame

|   | Country | GDP_Per_Capita |
|---|---------|----------------|
| 0 | USA     | 45000          |
| 1 | Canada  | 42000          |
| 2 | Germany | 52000          |
| 3 | UK      | 49000          |
| 4 | France  | 47000          |

# Plot a Bar Chart using Pandas

Step 3: Plot the DataFrame using Pandas



# Plot a Pie Chart using Pandas

A pie chart (or a circle chart) is a circular statistical graphic, which is divided into slices to illustrate numerical proportion.

Step 1: Prepare the data

|                 |     |
|-----------------|-----|
| Tasks Pending   | 300 |
| Tasks Ongoing   | 500 |
| Tasks Completed | 700 |

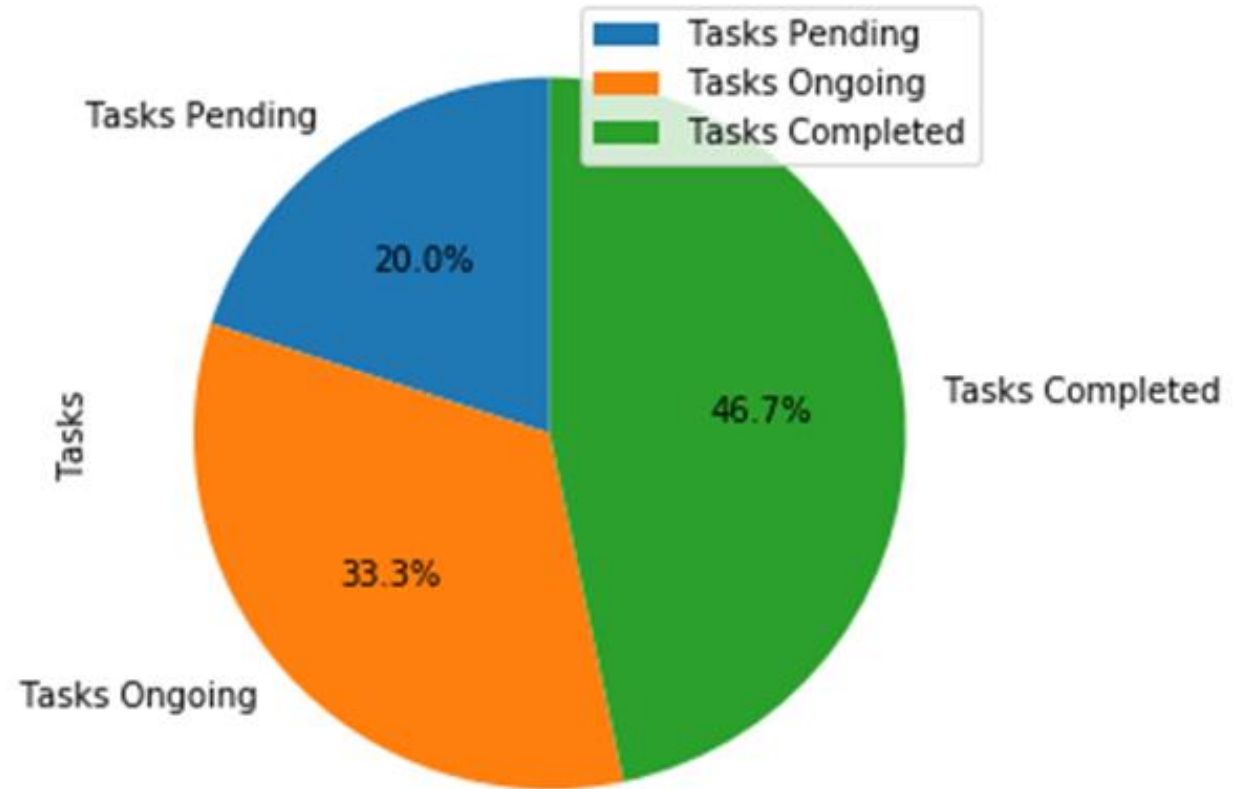
# Plot a Pie Chart using Pandas

Step 2: Create the DataFrame

|                 | Tasks |
|-----------------|-------|
| Tasks Pending   | 300   |
| Tasks Ongoing   | 500   |
| Tasks Completed | 700   |

# Plot a Pie Chart using Pandas

Step 3: Plot the DataFrame using Pandas





# Machine Learning



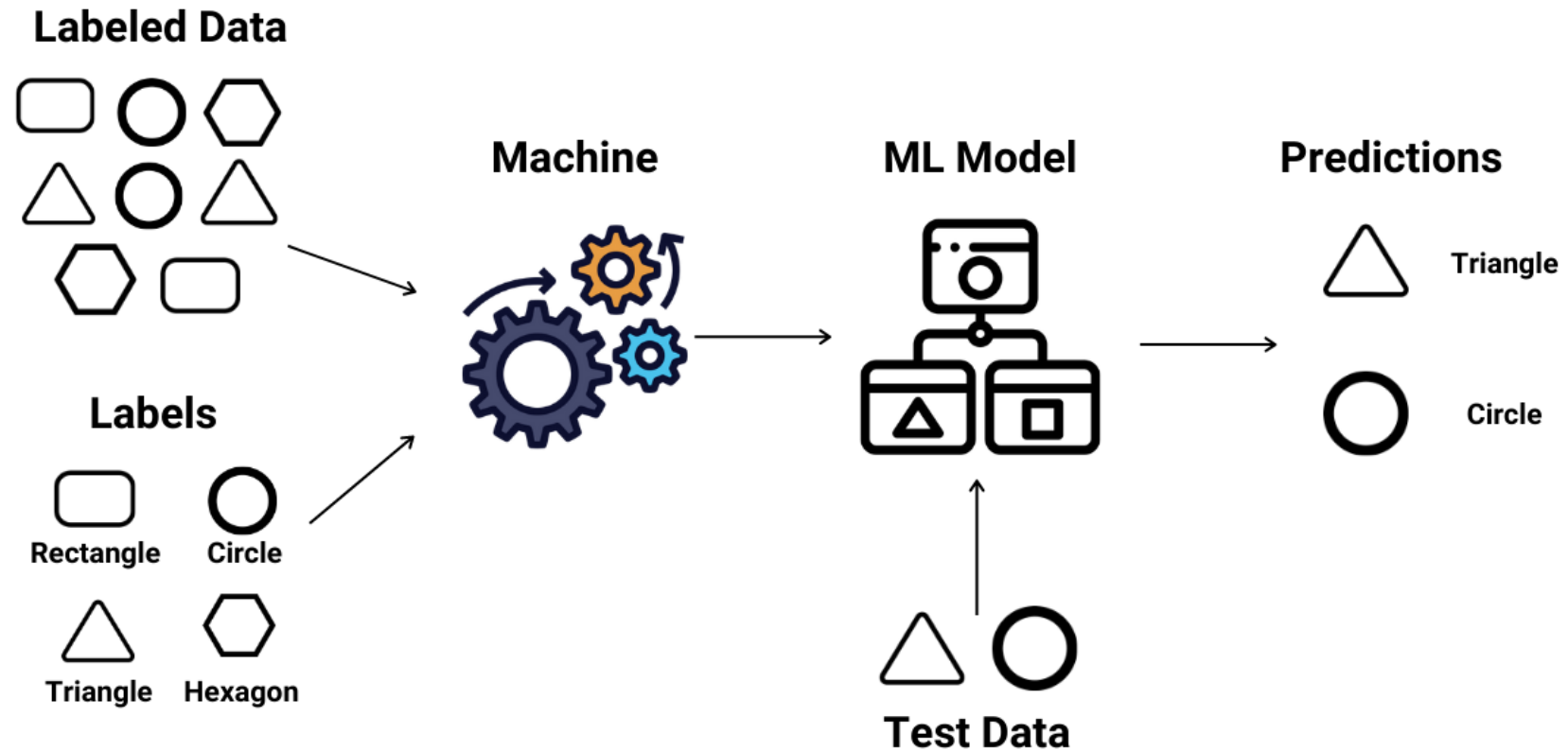
[https://miro.medium.com/max/629/1\\*\\_HoMKjrWahRil-JmwYW6zg.png](https://miro.medium.com/max/629/1*_HoMKjrWahRil-JmwYW6zg.png)

# List of Popular Dataset Websites

1. Data Government of India – <https://www.data.gov.in>
2. Earth Data - <https://earthdata.nasa.gov/>
3. Amazon and Microsoft Datasets, Azure and AWS  
<https://registry.opendata.aws/>  
<https://azure.microsoft.com/en-us/services/open-datasets/catalog>
4. Data World - <https://data.world/>
5. Lionbridge AI Datasets - <https://lionbridge.ai/datasets/>
6. UCI Machine Learning Repository -  
<https://archive.ics.uci.edu/ml/datasets.php>
7. Kaggle Datasets - <https://www.kaggle.com/datasets>



# Supervised Learning



[https://cdn-images-1.medium.com/max/1600/1\\*Iz7bCLrPTImnBDOOEyE3LA.png](https://cdn-images-1.medium.com/max/1600/1*Iz7bCLrPTImnBDOOEyE3LA.png)

# Regression vs Classification



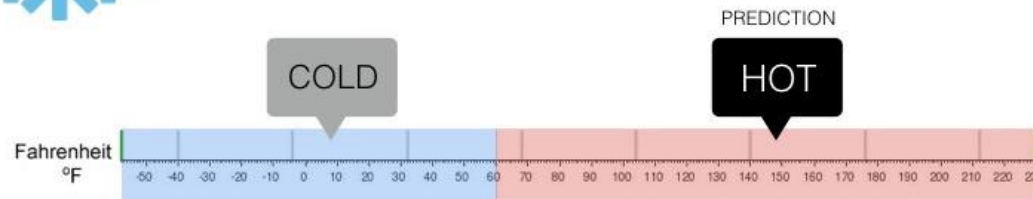
## Regression

What is the temperature going to be tomorrow?



## Classification

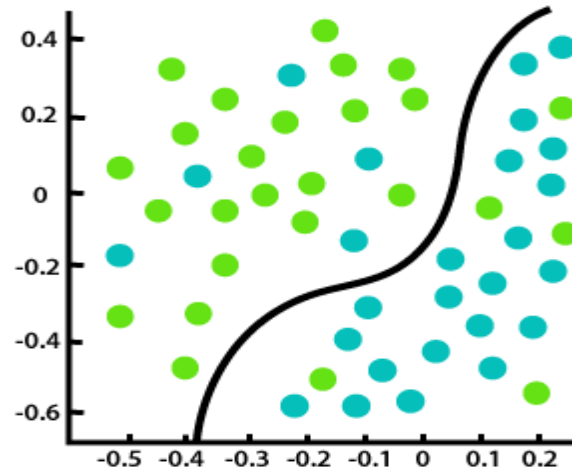
Will it be Cold or Hot tomorrow?



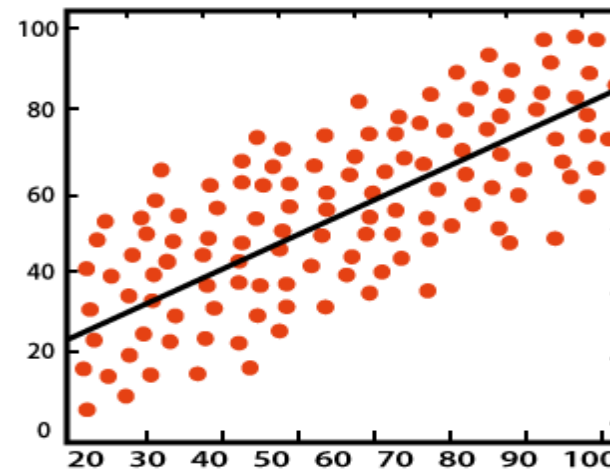
<https://www.javatpoint.com/regression-vs-classification-in-machine-learning>

# Regression vs Classification

- Regression algorithms are used to predict the continuous values.
- Classification algorithms are used to predict or Classify the discrete values.



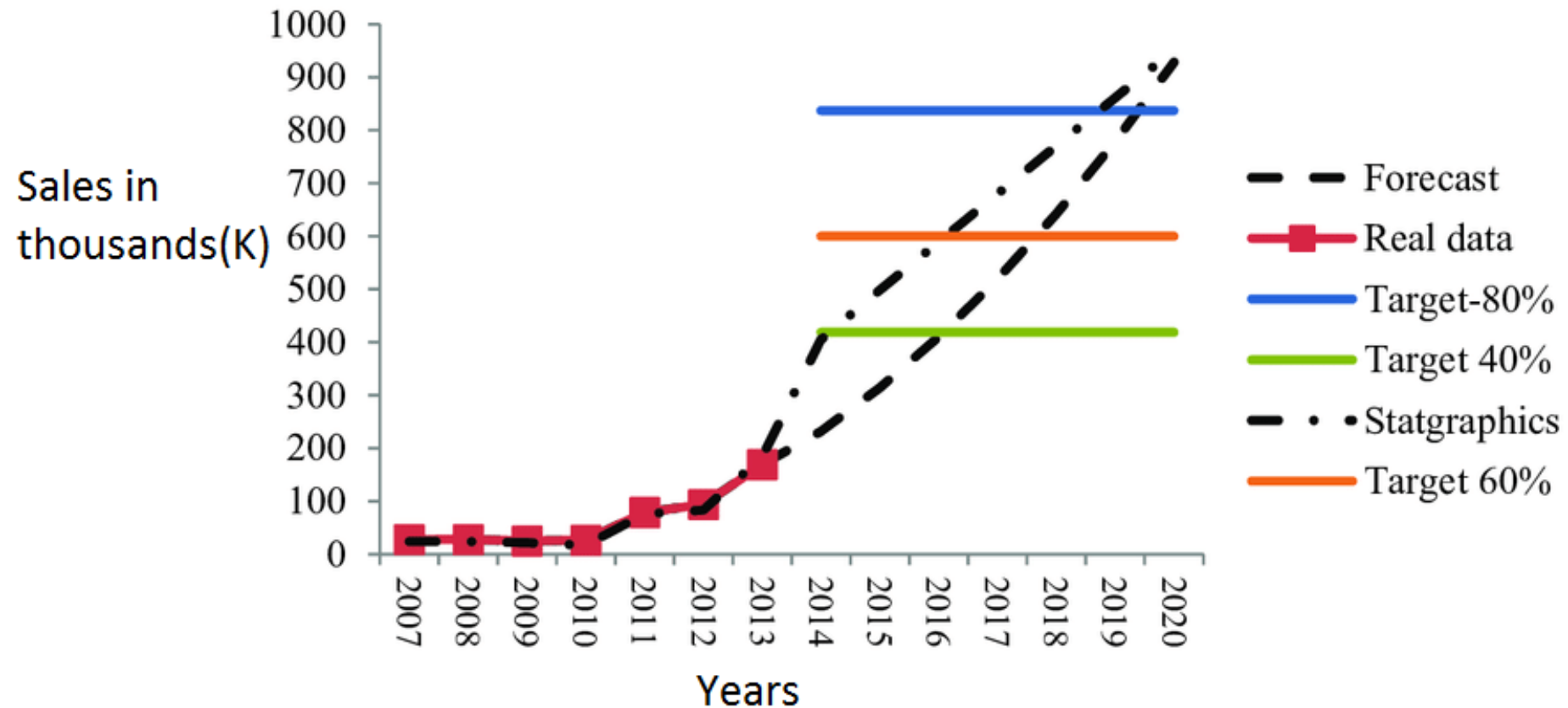
Classification



Regression

<https://www.javatpoint.com/regression-vs-classification-in-machine-learning>

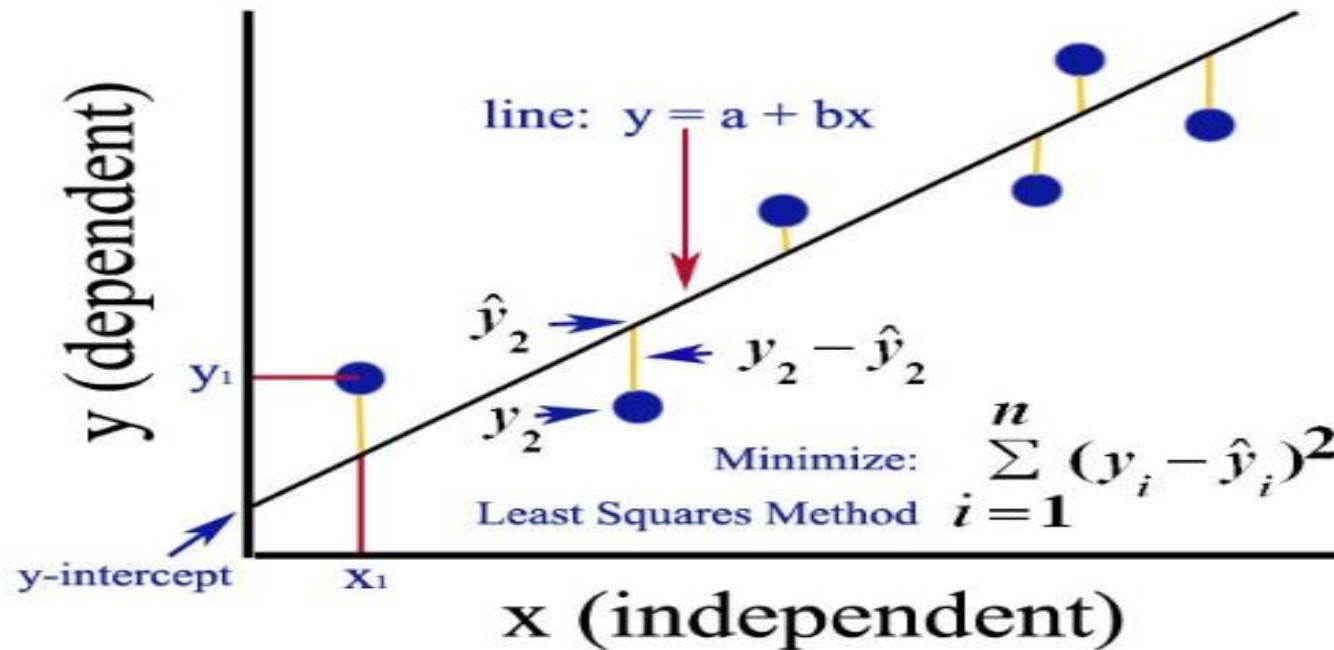
# Regression Analysis



<https://www.javatpoint.com/regression-vs-classification-in-machine-learning>

# Least Square Method

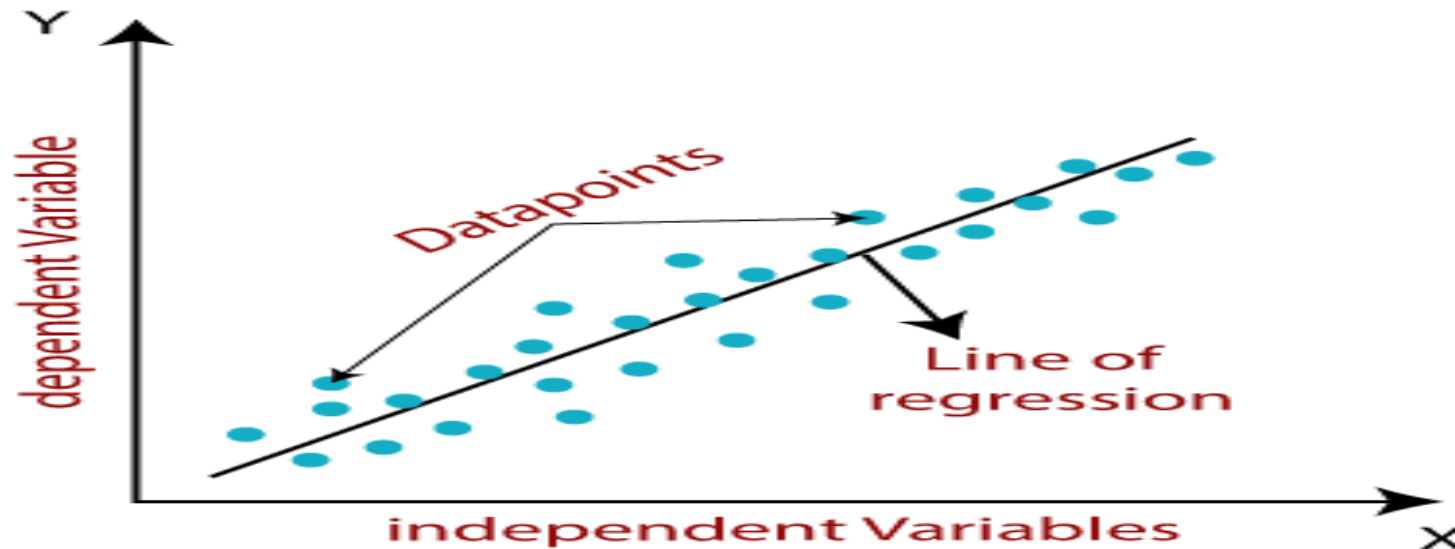
To find the best fit line that represents the relationship between an independent and dependent variable.



<https://medium.com/analytics-vidhya/ordinary-least-square-ols-method-for-linear-regression-ef8ca10aadfc>

# Linear Regression

It shows a linear relationship between a dependent (y) and one or more independent (x) variables.



<https://www.javatpoint.com/linear-regression-in-machine-learning>

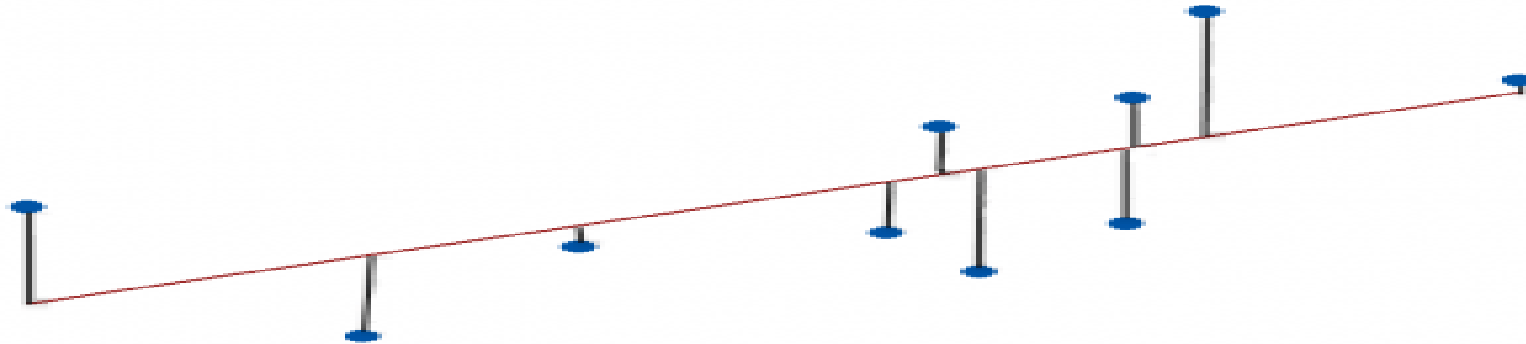


# Mathematical Intuition

- **Cost function:** It measures how a linear regression model is performing.
- **Gradient Descent:** To minimize the MSE by calculating the gradient of the cost function.
- **Model Performance:** Process of finding the best model out of various models.

# Ordinary Least Square Method

Estimates the parameters in a regression model by minimizing the sum of the squared residuals.



<https://statisticsbyjim.com/glossary/ordinary-least-squares/#:~:text=Ordinary%20least%20squares%2C%20or%20linear,and%20the%20corresponding%20fitted%20values>

# Linear Regression Implementation

- Import the Libraries

```
# importing the Linear Regression Model from Scikit Learn  
from sklearn.linear_model import LinearRegression
```

- Initialize our Linear Regression model

```
# initialize the Linear Regression model  
regression = LinearRegression()
```

- Fitting the Linear Regression Model

```
# fit the linear regression model  
regression.fit(X_train, y_train)
```

- Predict the test set Result

```
# Predict the Regression model  
y_pred = regression.predict(X_test)
```

# Linear Regression Evaluation Techniques

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}|$$

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y})^2$$

Where,

$\hat{y}$  – predicted value of  $y$

$\bar{y}$  – mean value of  $y$

<https://www.datatechnotes.com/2019/02/regression-model-accuracy-mae-mse-rmse.html>

# Azure Machine Learning **No Code Platform**

[https://docs.microsoft.com/en-us/azure/machine-learning/overview-what-is-azure-ml?WT.mc\\_id=aiml-0000-abornst](https://docs.microsoft.com/en-us/azure/machine-learning/overview-what-is-azure-ml?WT.mc_id=aiml-0000-abornst)

# Azure ML No Code Platform

A cloud service that allows building no-code machine learning models through a drag and drop visual interface.

Set of Azure  
Cloud Services



Python  
SDK

---

That enables  
you to:

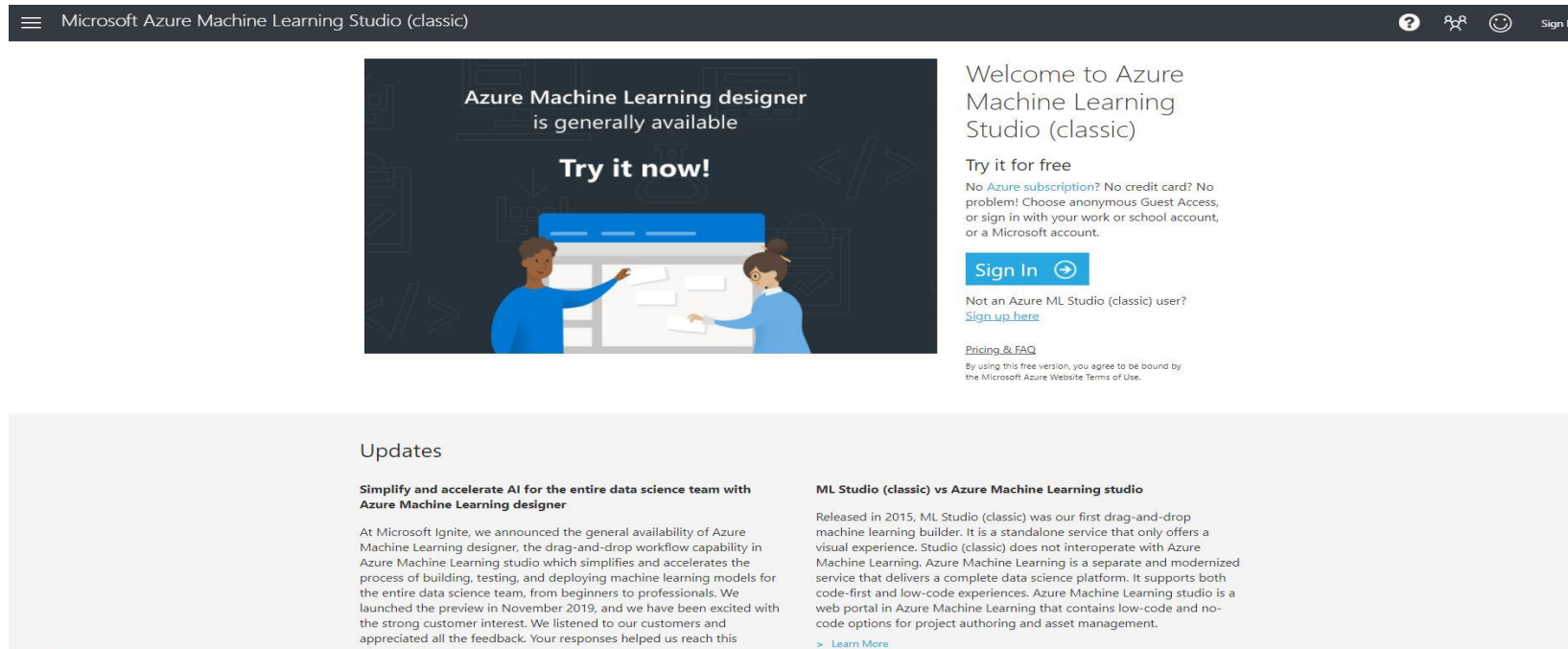
- ✓ Prepare Data
- ✓ Build Models
- ✓ Train Models

- ✓ Manage Models
- ✓ Track Experiments
- ✓ Deploy Models

[https://docs.microsoft.com/en-us/azure/machine-learning/overview-what-is-azure-ml?WT.mc\\_id=aiml-0000-abornst](https://docs.microsoft.com/en-us/azure/machine-learning/overview-what-is-azure-ml?WT.mc_id=aiml-0000-abornst)

# Practical - Azure ML Studio Briefing

Web-based integrated development environment (IDE) for developing data experiments.



The screenshot shows the Microsoft Azure Machine Learning Studio (classic) homepage. The header includes the Microsoft logo, the text "Microsoft Azure Machine Learning Studio (classic)", and navigation icons. The main content area features a large banner for "Azure Machine Learning designer" with a "Try it now!" button. To the right, there is a "Welcome to Azure Machine Learning Studio (classic)" section with a "Try it for free" button and a "Sign In" button. Below the banner, there is an "Updates" section with two articles: "Simplify and accelerate AI for the entire data science team with Azure Machine Learning designer" and "ML Studio (classic) vs Azure Machine Learning studio".

Microsoft Azure Machine Learning Studio (classic)

Azure Machine Learning designer is generally available

Try it now!

Welcome to Azure Machine Learning Studio (classic)

Try it for free

No Azure subscription? No credit card? No problem! Choose anonymous Guest Access, or sign in with your work or school account, or a Microsoft account.

Sign In

Not an Azure ML Studio (classic) user? [Sign up here](#)

Pricing & FAQ

By using this free version, you agree to be bound by the Microsoft Azure Website Terms of Use.

Updates

**Simplify and accelerate AI for the entire data science team with Azure Machine Learning designer**

At Microsoft Ignite, we announced the general availability of Azure Machine Learning designer, the drag-and-drop workflow capability in Azure Machine Learning studio which simplifies and accelerates the process of building, testing, and deploying machine learning models for the entire data science team, from beginners to professionals. We launched the preview in November 2019, and we have been excited with the strong customer interest. We listened to our customers and appreciated all the feedback. Your responses helped us reach this milestone. Thank you.

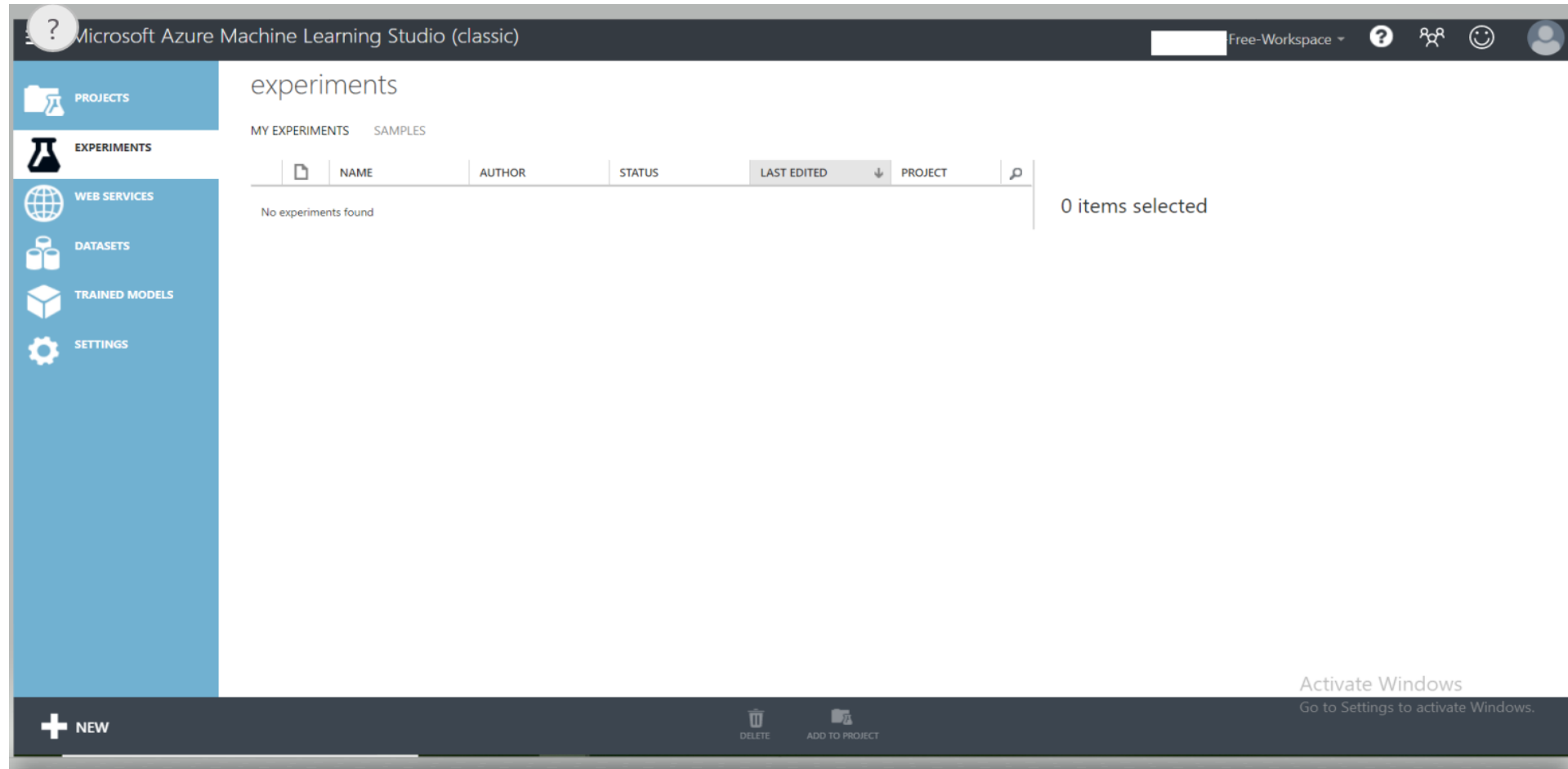
**ML Studio (classic) vs Azure Machine Learning studio**

Released in 2015, ML Studio (classic) was our first drag-and-drop machine learning builder. It is a standalone service that only offers a visual experience. Studio (classic) does not interoperate with Azure Machine Learning. Azure Machine Learning is a separate and modernized service that delivers a complete data science platform. It supports both code-first and low-code experiences. Azure Machine Learning studio is a web portal in Azure Machine Learning that contains low-code and no-code options for project authoring and asset management.

[Learn More](#)

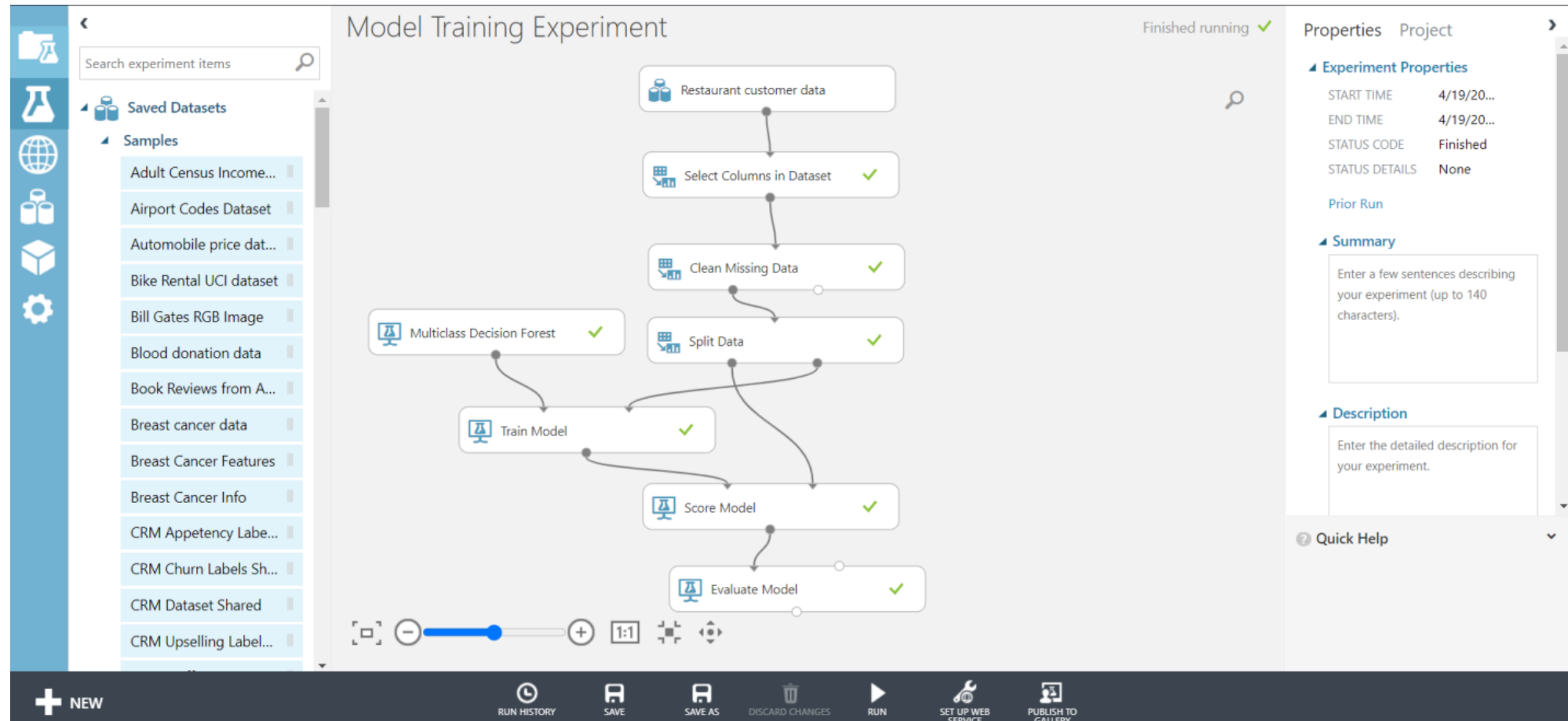
<https://studio.azureml.net/>

# Microsoft Azure workspace

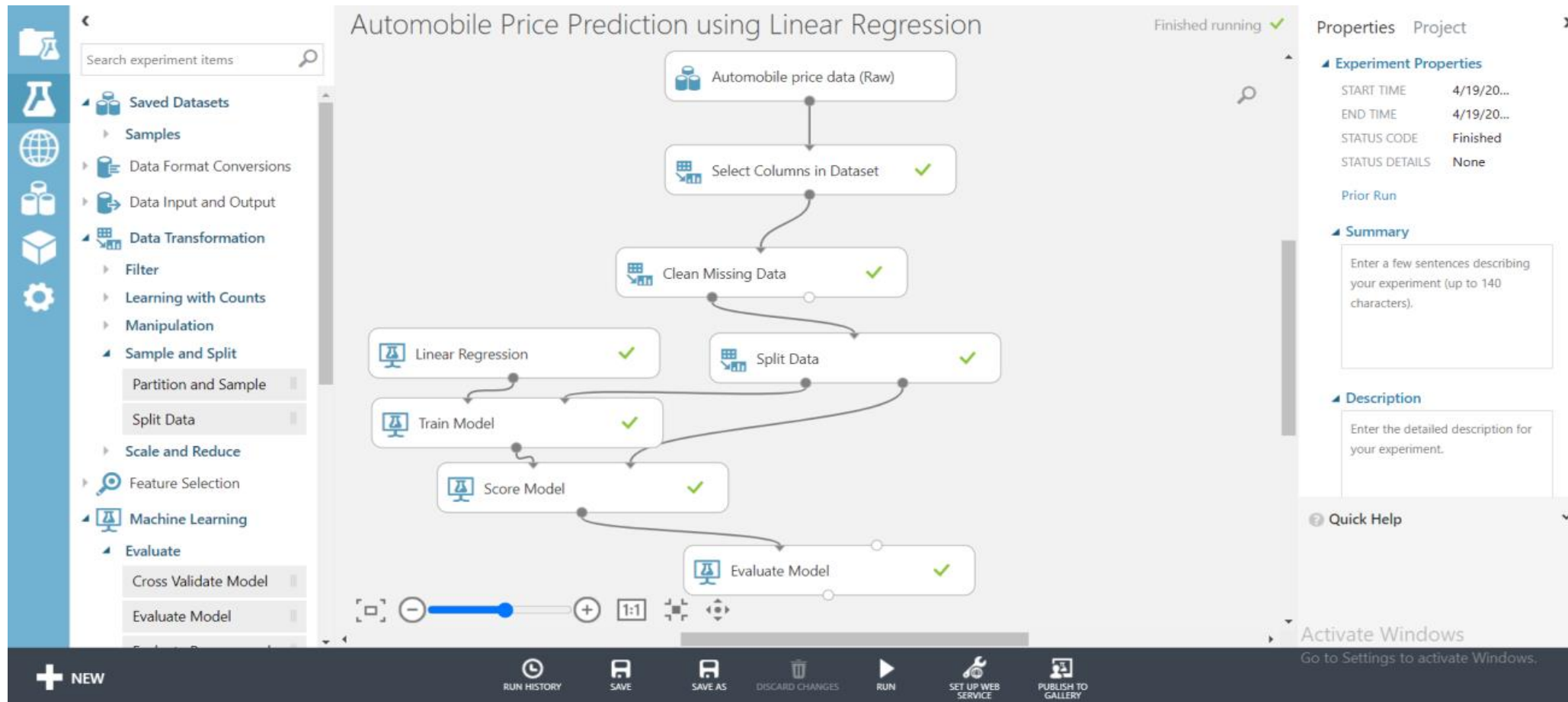




# Training a ML model



# Regression model with Azure ML Studio



# REFERENCES

1. [https://pandas.pydata.org/pandas-docs/stable/user\\_guide/io.html](https://pandas.pydata.org/pandas-docs/stable/user_guide/io.html)
2. <https://www.bigcommerce.com/ecommerce-answers/what-csv-file-and-what-does-it-mean-my-ecommerce-business/#:~:text=A%20CSV%20is%20a%20comma,Microsoft%20Excel%20or%20Google%20Spreadsheets.>
3. <https://fileinfo.com/extension/json>
4. <https://www.w3resource.com/JSON/structures.php>
5. <https://www.shanelynn.ie/python-pandas-read-csv-load-data-from-csv-files/>

# THANK YOU