ॐ

Project Report(Team HeHa)

(Women Empowerment in IT Field)

ES1101: COMPUTATIONAL AND DATA ANALYSIS

INSTITUTE OF ENGINEERING AND TECHNOLOGY(IET)
JK LAKSHMIPAT UNIVERSITY, JAIPUR

Prepared By –

Abhishek Saini (2021btech003)
Abhishek Swami (2021btech004)
Anurag Sharma (2021btech021)
Kapil Saini (2021btech056)

Section – **A**

SUBMITTED TO –

Dr. Umesh Gupta
Dr. Sonal Jain

INDEX

Abstract

The property Development Goals (SDGs), otherwise known as the world Goals, were taken on by the UN in 2015 as a widespread supply of inspiration to finish destitution, safeguard the planet, and guarantee that by 2030 all people appreciate harmony and success.

The seventeen SDGs are coordinated they understand that activity in one region can influence ends up in others which improvement ought to alter social, financial, and ecological maintainability. Nations have targeted on that specialize in progress for the folks that are uttermost behind. The SDGs are supposed to end the neediness, yearning, AIDS, and victimization of female and young women.

The imagination, expertise, innovation, and financial assets from all of society are vital to accomplish the SDGs in every specific situation.

Introduction

WHAT IS SUSTAINABE GOAL DEVELOPMENT?

The Sustainable Development Goals (SDGs), in any other case referred to as the Global Goals, had been taken on through the United Nations in 2015 as an all-inclusive supply of thought to cease neediness, steady the planet, and assure that through 2030 all people respect concord and success. The 17 SDGs are coordinated they understand that interest in a single place will affect outcomes in others and that development must modify social, monetary, and ecological maintainability.

SDG GOAL-5

Gender Equality

Orientation equity is not only a fundamental, not unusual place freedom, but an essential established order for a serene, wealthy, and viable world. There has been developed at some stage in the ultimate many years: More younger girls will school, much fewer young girls are restricted into early marriage, extra girls are serving in parliament and locations of authority, and guidelines are being modified to propel orientation equity. Despite those additions, many problems live: oppressive guidelines and regularly occurring practices live unavoidably, girls preserve on being underrepresented in any respect tiers of political authority, and 1 of each five girls and younger girls among a long time of 15 and forty-nine file encountering bodily or sexual brutality with the aid of using a private associate internal 12 months.

Team Target:

(Target 5.B)

"Enhance the use of enabling technology, in particular information and communications technology, to promote the empowerment of women."

For girls the sector over, statistics and correspondence innovations (ICT) may be applied for the man or woman security, higher admittance to training and occupations, economic incorporation or to get to essential clinical offerings statistics. Yet, advantages, for example, those rely on girls having extensive admittance to ICT which may be laboured with or forestalled with the aid of using some variables, such as reasonableness, essential substance, abilities, and security. SDG five intends to perform orientation equity and permit all girls and younger girls and requires upgraded usage of empowering innovation - ICTs specifically - to enhance the strengthening of girls.

To help with reworking obligations proper into it, the International Chamber of Commerce has collaborated with UN Women - the global boss for orientation equity - to have a facet event throughout the HLPF. The event, entitled Accelerating Women's Economic Empowerment to Achieve the 2030 Agenda will show off the global endeavours companions have left on to hold girls' economic strengthening to the very front of all of the SDG targets. Through advancement, venture, and the development of gadgets and administrations, the non-public region assumes an extensive element in propelling orientation equity and running at the lifestyles of girls. While girls make up over 1/2 of the full populace, they likewise cope with 70% of the sector's poor. As indicated with the aid of using research, girls reinvest 80% of every greenback made yet again into their family, implying that pragmatic assistance for the economic strengthening of girls is a crucial degree closer to killing neediness and advancing success.

Objectives-

1. To analyse and study IT field expertise in Machine learning and Data science.

2. To analyse and study, the status of the jobs in the IT field.

3. To investigate every one of the ventures and compensations in the tech field and look at all things.

4. To explore how various factors affects in mobile cellular.

Problems Statements-

- To analyse and study IT field expertise in Machine learning and Data science.

1.1 To breakdown the average experience by the individual

1.2 To analyse people based on knowledge in the tech field and differentiate with genders.

1.3 To investigate women strengthening as per nations.

- **To analyse and study, the status of the jobs in the IT field.**

2.1 To examine the gender differences in technical employment throughout the world.

2.2 On the basis of previously existing data, assess and forecast the future of India's job statuses.

2.3 To analyse and look at the representative's age in the Tech Industry with male and female.

- **To investigate every one of the investment and compensations in the tech field.**

3.1 To concentrate on the compensations given by organizations and see the inclinations done by organizations.

3.2 To determine the number of individuals in the tech area and perceive how much women is putting resources into it.

- To explore how various factors affects in mobile cellular.

4.1 To relate the Proportion of female and male population using the Internet around the world. And to check the claim "***Female users use less mobile phone than male***"

4.2 To predict total number cellular data user for next decade based on current data.

Literature Review

1.

There is a demanding and ongoing loss of girls hired in Artificial Intelligence (AI) and information technology fields. According to the World Economic Forum, girls make-up simplest 26% of the staff in information and AI roles international. In 2012, the OECD surveyed 15-year-antique UK college students located that 41% of female agreed with the `I`m now no longer correct at math` statement, whilst simplest 24% of boys agreed. In 2015, the OECD surveyed 15-year-olds and located that 4.6% of boys had been predicted to paintings as IT professionals via way of means of age 30, whilst simplest 0.5% of female predicted the equal from them. Women in information and AI are much less withinside the enterprise historically consists of extra technical skills, for example, the Technology / IT zone in addition to holds fewer technical skills (e.g., DevOps). In addition, there are fewer girls than guys in C-suite positions in maximum regions of

industries, and that is even greater marked via way of means of information and AI sports withinside the generation zone.

2.

Around the world, locating a process is tons harder for girls than it's far for guys. When girls are hired, they generally tend to paintings in low-first-rate jobs in inclined conditions, and there may be little development forecast shortly. Explore this Info Story to get the information at the back of the traits and research greater approximately the exclusive boundaries maintaining girls returned from first rate paintings. Women`s participation in IT is better each in evaluation to different sectors withinside the U.S.A . and evaluation to illustration withinside the zone in different international locations. The boom in numbers of girls in engineering training and the mushrooming of numerous non-public engineering schools for the reason that Nineties catered to a boom in call for engineers withinside the IT enterprise. Women who need to paintings have a tougher time locating a process than guys. This trouble is mainly marked in Northern Africa and the Arab States, in which

unemployment fees for girls exceed 16%. So on this report, we're going to see how tons distinction in intercourse ratio in jobs at precise IT zone via way of means of our Objective and Problem Statements. we're going to examine the information and observe it. the information that we've got is the survey of Kaggle.

3.

The gender salary hole refers back to the distinction in income among girls and guys. Experts have calculated this hole in more than one ways, however the various calculations factor to a consensus: Women continuously earn much less than guys, and the distance is wider for maximum girls of colour. As indicated via way of means of the maximum latest World Economic Forum`s (WEF) Global Gender Gap Report 2018, India placed 108th out of 149 countries at the orientation hollow file. The international rundown turned into crowned via way of means of Iceland for the 10th non-stop year, having close over 85.8% of its standard orientation hollow. It is likewise withinside the gender salary hole we're going to see how tons distinction in salaries and appearance as much as in India additionally. We

all should see while this discrimination goes to and I'll percentage the information of the survey of Kaggle in which we are able to see this and affirm our goal and trouble statement. Women make up round 250 million fewer on line customers than guys, and the disparity is widening (from eleven percentage in 2013 to twelve percentage in 2016).

4.

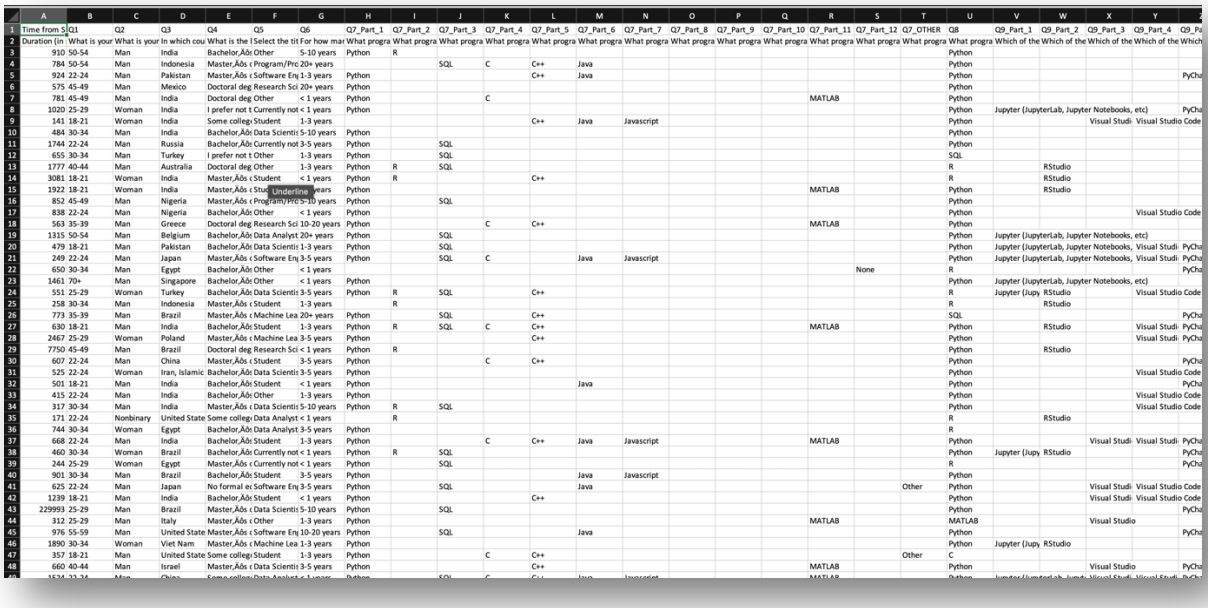
In the modern-day period, get entry to is concentrated; globally, 53% of the population (3.nine billion humans) stays disconnected, and simply one in each ten humans in numerous of Africa's poorer and additionally maximum volatile international locations has an Internet connection. Increased get entry to on line sources for girls and female is vital to making sure that they do now no longer fall at the back of in an increasing number of virtual world, and it may, in positive cases, enhance girl’s hobby withinside the opportunities supplied via way of means of generation and ICTs. Women`s get entry to cell net maintains to boom throughout low- and middle-earnings international locations with 112 million extra lady customers getting on line in 2020. Despite this, the gender hole stays substantial. Women are

7% much less probably than guys to personal a cell telecellsmartphone and 15% much less probably to apply cell net. There are nevertheless 234 million fewer girls than guys gaining access to cell net. With the COVID-19 pandemic evolving throughout the world, there has in no way been a greater pressing time to deal with this issue. The Mobile Gender Gap Report 2021 highlights how the cell gender hole maintains to enhance in South Asia, however much less so in different regions. It explores the important thing boundaries stopping girl's same get entry to cell net in addition to the upward thrust of girl's telecellsmartphone ownership, substantially in India.

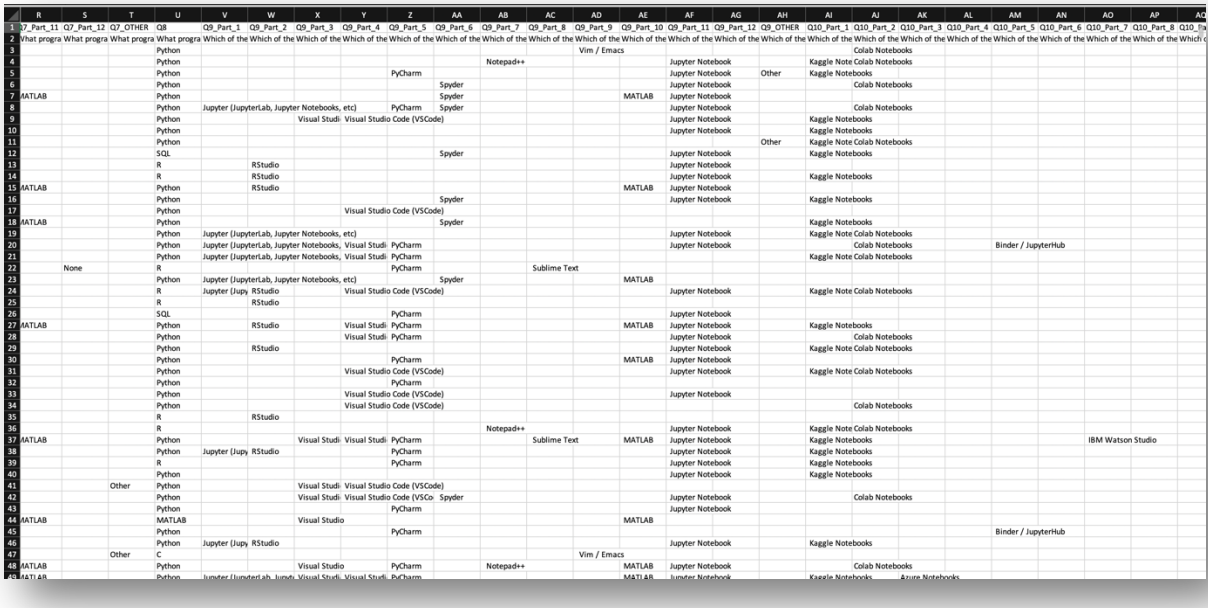
Data Collection:

We choose our team Goal "Gender Equality" from SDG and after reading some articles on gender Equality which is related to technology so we stuck on the technology part and when we have to select the target we choose Target 5.B.

After this, we started to make our team objectives and start searching data for it, after searching on google a lot we found our first dataset on gender equality on Kaggle, which is a Kaggle survey which is around $ 30,000 so we choose that and make our problem statement on it.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	Time from S	Q1	Q2	Q3	Q4	Q5	Q6	Q7_Part_1	Q7_Part_2	Q7_Part_3	Q7_Part_4	Q7_Part_5	Q7_Part_6	Q7_Part_7	Q7_Part_8	Q7_Part_9	Q7_Part_10	Q7_Part_11	Q7_Part_12	Q7_OTHER	Q8	Q9_Part_1	Q9_Part_2	Q9_Part_3	Q9_Part_4	Q9_Pa
2	Duration (in	What is your	What is your	In which cou	What is the l	Select the tit	For how mar	What progra	What progra	What progra	What progra	What progra	What progra	What progra	What progra	What progra	What progra	What progra	What progra	What progra	What progra	Which of the	Which of the	Which of the	Which of the	Which
3	910	50-54	Man	India	Bachelor,Âôs	Other	5-10 years	Python	R												Python					
4	784	50-54	Man	Indonesia	Master,Âôs	Program/Prc	20+ years			SQL	C	C++	Java								Python					
5	924	22-24	Man	Pakistan	Master,Âôs	Software Eng	1-3 years	Python				C++	Java								Python					PyCha
6	575	45-49	Man	Mexico	Doctoral deg	Research Sci	20+ years	Python													Python					
7	781	45-49	Man	India	Doctoral deg	Other	< 1 years	Python			C							MATLAB			Python					
8	1020	25-29	Woman	India	I prefer not t	Currently not	< 1 years	Python													Python	Jupyter (JupyterLab, Jupyter Notebooks, etc)				PyCha
9	141	18-21	Woman	India	Some college	Student	1-3 years					C++	Java	Javascript							Python			Visual Studi	Visual Studio	Code
10	484	30-34	Man	India	Bachelor,Âôs	Data Scientis	5-10 years	Python													Python					
11	1744	22-24	Man	Russia	Bachelor,Âôs	Currently not	3-5 years	Python		SQL											Python					
12	655	30-34	Man	Turkey	I prefer not t	Other	1-3 years	Python		SQL											SQL					
13	1777	40-44	Man	Australia	Doctoral deg	Other	1-3 years	Python	R	SQL											R		RStudio			
14	3081	18-21	Woman	India	Master,Âôs	Student	< 1 years	Python	R			C++									R		RStudio			
15	1922	18-21	Woman	India	Master,Âôs	Student	1-3 years	Python										MATLAB			Python		RStudio			
16	852	45-49	Man	Nigeria	Master,Âôs	Program/Prc	5-10 years	Python		SQL											Python					
17	838	22-24	Man	Nigeria	Bachelor,Âôs	Other	< 1 years	Python													Python				Visual Studio	Code
18	563	35-39	Man	Greece	Doctoral deg	Research Sci	10-20 years	Python			C	C++						MATLAB			Python					
19	1315	50-54	Man	Belgium	Bachelor,Âôs	Data Analyst	20+ years	Python		SQL											Python	Jupyter (JupyterLab, Jupyter Notebooks, etc)				
20	479	18-21	Man	Pakistan	Bachelor,Âôs	Data Scientis	1-3 years	Python		SQL											Python	Jupyter (JupyterLab, Jupyter Notebooks,		Visual Studi		PyCha
21	249	22-24	Man	Japan	Master,Âôs	Software Eng	3-5 years	Python		SQL	C		Java	Javascript							Python	Jupyter (JupyterLab, Jupyter Notebooks,		Visual Studi		PyCha
22	650	30-34	Man	Egypt	Bachelor,Âôs	Other	< 1 years												None		R					PyCha
23	1461	70+	Man	Singapore	Bachelor,Âôs	Other	< 1 years	Python													Python	Jupyter (JupyterLab, Jupyter Notebooks, etc)				
24	551	25-29	Woman	Turkey	Bachelor,Âôs	Data Scientis	3-5 years	Python	R	SQL		C++									R	Jupyter (Jupy	RStudio		Visual Studio	Code
25	258	30-34	Man	Indonesia	Master,Âôs	Student	1-3 years		R												R		RStudio			
26	773	35-39	Man	Brazil	Master,Âôs	Machine Lea	20+ years	Python		SQL		C++									SQL					PyCha
27	630	18-21	Man	India	Bachelor,Âôs	Student	1-3 years	Python	R	SQL	C	C++						MATLAB			Python		RStudio		Visual Studi	PyCha
28	2467	25-29	Woman	Poland	Master,Âôs	Machine Lea	3-5 years	Python				C++									Python				Visual Studi	PyCha
29	7750	45-49	Man	Brazil	Doctoral deg	Research Sci	< 1 years	Python	R												Python		RStudio			
30	607	22-24	Man	China	Master,Âôs	Student	3-5 years	Python			C	C++									Python					PyCha
31	525	22-24	Woman	Iran, Islamic	Bachelor,Âôs	Data Scientis	3-5 years	Python													Python				Visual Studio	Code
32	501	18-21	Man	India	Bachelor,Âôs	Student	< 1 years	Python					Java								Python					PyCha
33	415	22-24	Man	India	Bachelor,Âôs	Other	1-3 years	Python													Python				Visual Studio	Code
34	317	30-34	Man	India	Master,Âôs	Data Scientis	5-10 years	Python	R	SQL											Python				Visual Studio	Code
35	171	22-24	Nonbinary	United State	Some college	Data Analyst	< 1 years		R												R		RStudio			
36	744	30-34	Woman	Egypt	Bachelor,Âôs	Data Analyst	3-5 years	Python													R					
37	668	22-24	Man	India	Bachelor,Âôs	Student	1-3 years	Python			C	C++	Java	Javascript				MATLAB			Python			Visual Studi	Visual Studi	PyCha
38	460	30-34	Woman	Brazil	Bachelor,Âôs	Currently not	< 1 years	Python	R	SQL											Python	Jupyter (Jupy	RStudio			PyCha
39	244	25-29	Woman	Egypt	Master,Âôs	Currently not	< 1 years	Python		SQL											R					PyCha
40	901	30-34	Man	Brazil	Bachelor,Âôs	Student	3-5 years	Python					Java	Javascript							Python					
41	625	22-24	Man	Japan	No formal ec	Software Eng	3-5 years	Python		SQL			Java							Other	Python			Visual Studi	Visual Studio	Code
42	1239	18-21	Man	India	Bachelor,Âôs	Student	< 1 years	Python				C++									Python			Visual Studi	Visual Studio	Code
43	229993	25-29	Man	Brazil	Master,Âôs	Data Scientis	5-10 years	Python		SQL											Python					PyCha
44	312	25-29	Man	Italy	Master,Âôs	Other	1-3 years	Python										MATLAB			MATLAB			Visual Studio		
45	976	55-59	Man	United State	Master,Âôs	Software Eng	10-20 years	Python		SQL			Java								Python					PyCha
46	1890	30-34	Woman	Viet Nam	Master,Âôs	Machine Lea	1-3 years	Python													Python	Jupyter (Jupy	RStudio			
47	357	18-21	Man	United State	Some college	Student	1-3 years	Python			C	C++								Other	C					
48	660	40-44	Man	Israel	Master,Âôs	Data Scientis	3-5 years	Python				C++						MATLAB			Python			Visual Studio		PyCha



	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN	AO	AP	AQ
1	7_Part_11	Q7_Part_12	Q7_OTHER	Q8	Q9_Part_1	Q9_Part_2	Q9_Part_3	Q9_Part_4	Q9_Part_5	Q9_Part_6	Q9_Part_7	Q9_Part_8	Q9_Part_9	Q9_Part_10	Q9_Part_11	Q9_Part_12	Q9_OTHER	Q10_Part_1	Q10_Part_2	Q10_Part_3	Q10_Part_4	Q10_Part_5	Q10_Part_6	Q10_Part_7	Q10_Part_8	Q10_Fa
2	What progra	What progra	What progra	What progra	Which of the	Which of the	Which of the	Which of the	Which of the	Which of the	Which of the	Which of the	Which of the	Which of the	Which of the	Which of the	Which of the	Which of the	Which of the	Which of the	Which of the	Which of the	Which of the	Which of the	Which of the	Which
3				Python									Vim / Emacs						Colab Notebooks							
4				Python							Notepad++				Jupyter Notebook			Kaggle Note	Colab Notebooks							
5				Python					PyCharm						Jupyter Notebook		Other	Kaggle Note	Colab Notebooks							
6				Python						Spyder					Jupyter Notebook				Colab Notebooks							
7	MATLAB			Python						Spyder				MATLAB	Jupyter Notebook											
8				Python	Jupyter (JupyterLab, Jupyter Notebooks, etc)				PyCharm	Spyder					Jupyter Notebook				Colab Notebooks							
9				Python			Visual Studi	Visual Studio Code (VSCode)							Jupyter Notebook			Kaggle Note	Colab Notebooks							
10				Python											Jupyter Notebook			Kaggle Note	Colab Notebooks							
11				Python													Other	Kaggle Note	Colab Notebooks							
12				SQL						Spyder					Jupyter Notebook			Kaggle Note	Colab Notebooks							
13				R		RStudio									Jupyter Notebook											
14				R		RStudio									Jupyter Notebook			Kaggle Note	Colab Notebooks							
15	MATLAB			Python		RStudio								MATLAB	Jupyter Notebook											
16				Python						Spyder					Jupyter Notebook			Kaggle Note	Colab Notebooks							
17				Python				Visual Studio Code (VSCode)																		
18	MATLAB			Python						Spyder								Kaggle Note	Colab Notebooks							
19				Python	Jupyter (JupyterLab, Jupyter Notebooks, etc)										Jupyter Notebook			Kaggle Note	Colab Notebooks							
20				Python	Jupyter (JupyterLab, Jupyter Notebooks,			Visual Studi	PyCharm						Jupyter Notebook				Colab Notebooks			Binder / JupyterHub				
21				Python	Jupyter (JupyterLab, Jupyter Notebooks,			Visual Studi	PyCharm									Kaggle Note	Colab Notebooks							
22		None		R					PyCharm			Sublime Text														
23				Python	Jupyter (JupyterLab, Jupyter Notebooks, etc)					Spyder				MATLAB													
24				R	Jupyter (Jupy	RStudio		Visual Studio Code (VSCode)							Jupyter Notebook			Kaggle Note	Colab Notebooks							
25				R		RStudio																				
26				SQL					PyCharm						Jupyter Notebook											
27	MATLAB			Python		RStudio		Visual Studi	PyCharm					MATLAB	Jupyter Notebook			Kaggle Note	Colab Notebooks							
28				Python				Visual Studi	PyCharm						Jupyter Notebook				Colab Notebooks							
29				Python		RStudio									Jupyter Notebook			Kaggle Note	Colab Notebooks							
30				Python					PyCharm					MATLAB	Jupyter Notebook											
31				Python				Visual Studio Code (VSCode)							Jupyter Notebook			Kaggle Note	Colab Notebooks							
32				Python					PyCharm																	
33				Python				Visual Studio Code (VSCode)							Jupyter Notebook											
34				Python				Visual Studio Code (VSCode)											Colab Notebooks							
35				R		RStudio																				
36				R							Notepad++				Jupyter Notebook			Kaggle Note	Colab Notebooks							
37	MATLAB			Python			Visual Studi	Visual Studi	PyCharm			Sublime Text		MATLAB	Jupyter Notebook			Kaggle Note	Colab Notebooks					IBM Watson Studio		
38				Python	Jupyter (Jupy	RStudio			PyCharm						Jupyter Notebook			Kaggle Note	Colab Notebooks							
39				R					PyCharm						Jupyter Notebook			Kaggle Note	Colab Notebooks							
40				Python											Jupyter Notebook			Kaggle Note	Colab Notebooks							
41			Other	Python			Visual Studi	Visual Studio Code (VSCode)																		
42				Python			Visual Studi	Visual Studio Code (VSCo		Spyder					Jupyter Notebook				Colab Notebooks							
43				Python					PyCharm						Jupyter Notebook											
44	MATLAB			MATLAB			Visual Studio							MATLAB												
45				Python					PyCharm													Binder / JupyterHub				
46				Python	Jupyter (Jupy	RStudio									Jupyter Notebook			Kaggle Note	Colab Notebooks							
47			Other	C									Vim / Emacs													
48	MATLAB			Python			Visual Studio		PyCharm		Notepad++			MATLAB	Jupyter Notebook				Colab Notebooks							

So this is existing RAW dataset, we have to clean and sort it and it has around more than 25,000 rows and more than 300 columns.

So after sorting and cleaning of dataset we actually found some difficulties in our data, we use Python to sort it and clean it and after a lot of struggle we have first dataset.

Modules Import

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
```

Const And Global Variables

```
RAW_DATASET_LOC = "Obj2_Dataset/kaggle_survey_2021_responses.csv"
```

Data Cleaning

```
raw_dataset_df = pd.read_csv(RAW_DATASET_LOC, low_memory=False, skiprows=[0])
raw_dataset_df = raw_dataset_df.iloc[:, :232]
raw_dataset_df.head()
```

[11]:
```
dataset = pd.read_csv("Obj2_Dataset/New_Ra-Do-Woo's Dataset.csv")
dataset.drop("Unnamed: 0", inplace=True, axis=1)
dataset.head()
```

[11]:

	Response Time	Age	Gender	Country	Education	Job Title	Experience	Python	R	SQL	...	Neptune.ai	Weights & Biases	Comet.ml	Sacred + Omniboard	TensorBoard	Guild.ai	Polyaxon	ClearML	Domino Model Monitor	MLflow
0	910	50-54	Man	India	Bachelor's degree	Other	5-10 years	1	1	0	...	0	0	0	0	0	0	0	0	0	0
1	784	50-54	Man	Indonesia	Master's degree	Program/Project Manager	20+ years	0	0	1	...	0	0	0	0	0	0	0	0	0	0
2	924	22-24	Man	Pakistan	Master's degree	Software Engineer	1-3 years	1	0	0	...	0	0	0	0	0	0	0	0	0	0
3	575	45-49	Man	Mexico	Doctoral degree	Research Scientist	20+ years	1	0	0	...	0	0	0	0	0	0	0	0	0	0
4	781	45-49	Man	India	Doctoral degree	Other	< 1 years	1	0	0	...	0	1	0	0	0	0	0	0	0	0

5 rows × 169 columns

[26]:
```
dataset = dataset.loc[:, "Python":]
dataset_matrix = dataset.to_numpy()
# dataset_matrix.reshape(2,2)
```

Normalizing Dataset

[10]:
```
raw_dataset_df_1 = raw_dataset_df.iloc[:, :7]
raw_dataset_df_2 = raw_dataset_df.iloc[:, 7:]
raw_dataset_df_2_cols = raw_dataset_df_2.columns
raw_dataset_df_2.fillna(0, inplace=True)
raw_dataset_df_2 = (raw_dataset_df_2[raw_dataset_df_2_cols] != 0).astype(int)
```

[11]:
```
dataset = pd.concat([raw_dataset_df_1, raw_dataset_df_2], axis=1)
dataset.to_csv("Obj2_Dataset/New_Ra-Do-Woo's Dataset.csv")
```

[12]:
```
gender = dataset['Gender']
experience = dataset['Experience']
ps_one_df = pd.concat([gender, experience], axis=1)
ps_one_df.groupby("Experience").sum()
```

[12]:

	Gender
Experience	
1-3 years	ManWomanManManManManManManManManWomanManWomanM...
10-20 years	ManManManManManManManManManManManManManManM...
20+ years	ManManManManManManManManManManManWomanManMa...
3-5 years	ManManWomanWomanManWomanWomanManManManManMa...
5-10 years	ManManManManManWomanWomanWomanManManManManM...
< 1 years	ManWomanWomanWomanManManManManNonbinaryWoma...
I have never written code	ManManManManWomanWomanWomanWomanManWomanManWom...

[13]:
```
skill_set = dataset.iloc[:, 7:].sum(axis=1)
gender = dataset['Gender']
ps_one_df = pd.concat([gender, skill_set], axis=1)
ps_one_df.columns = ["Gender", "Skill Set"]
ps_one_df = ps_one_df.groupby("Gender").sum()
ps_one_df.sort_values("Skill Set", ascending=False, inplace=True)
ps_one_df
```

[13]:

	Skill Set
Gender	
Man	332690
Woman	63212
Prefer not to say	5946
Nonbinary	1295

```
columns_to_drop = [
    'None',
    'MATLAB',
    'Alteryx',
    'Other',
    '3, etc)',
    'No / None',
    'Approximately how many times have you used a TPU (tensor processing unit)?',
    'For how many years have you used machine learning methods?',
    'CNN, etc)',
    '3, BERT, XLnet, etc)',
    'Selected Choice',
    'What is the size of the company where you are employed?',
    'Approximately how many individuals are responsible for data science workloads at your place of business?',
    'Does your current employer incorporate machine learning methods into their business?',
    'Analyze and understand data to influence product or business decisions',
    'Build and/or run the data infrastructure that my business uses for storing, analyzing, and operationalizing data',
    'Build prototypes to explore applying machine learning to new areas',
    'Build and/or run a machine learning service that operationally improves my product or workflows',
    'Experimentation and iteration to improve existing ML models',
    'Do research that advances the state of the art of machine learning',
    'None of these activities are an important part of my role at work',
    'What is your current yearly compensation (approximate $USD)?',
    'Approximately how much money have you (or your team) spent on machine learning and/or cloud computing services at home (or at work) in the past 5 years (approximate $USD)?',
]
```

Stage 2 - Drop Unreliable Columns

```
raw_dataset_df.drop(columns_to_drop, axis=1, inplace=True)
raw_dataset_df
```

	Response Time	Age	Gender	Country	Education	Job Title	Experience	Python	R	SQL	...	Neptune.ai	Weights & Biases	Comet.ml	Sacred + Omniboard	TensorBoard	Guild.ai	Polyaxon	ClearML	Domino Model Monitor	MLflow
0	910	50-54	Man	India	Bachelor's degree	Other	5-10 years	Python	R	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	784	50-54	Man	Indonesia	Master's degree	Program/Project Manager	20+ years	NaN	NaN	SQL	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	924	22-24	Man	Pakistan	Master's degree	Software Engineer	1-3 years	Python	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Stage 1 - Simplifying Columns Name

```
simplified_columns = {i:i.split("-")[-1].strip() for i in raw_dataset_df.iloc[:, 7:].columns}
raw_dataset_df.rename(columns=simplified_columns, inplace=True)
raw_dataset_df
```

	Response Time	Age	Gender	Country	Education	Job Title	Experience	Python	R	SQL	...	Weights & Biases	Comet.ml	Sacred + Omniboard	TensorBoard	Guild.ai	Polyaxon	ClearML	Domino Model Monitor	MLflow	No / None
0	910	50-54	Man	India	Bachelor's degree	Other	5-10 years	Python	R	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	No / None
1	784	50-54	Man	Indonesia	Master's degree	Program/Project Manager	20+ years	NaN	NaN	SQL	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	924	22-24	Man	Pakistan	Master's degree	Software Engineer	1-3 years	Python	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	No / None
3	575	45-49	Man	Mexico	Doctoral degree	Research Scientist	20+ years	Python	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	781	45-49	Man	India	Doctoral degree	Other	< 1 years	Python	NaN	NaN	...	Weights & Biases	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
25968	1756	30-34	Man	Egypt	Bachelor's degree	Data Analyst	1-3 years	Python	NaN	SQL	...	NaN	NaN	NaN	TensorBoard	NaN	NaN	NaN	NaN	NaN	NaN
25969	253	22-24	Man	China	Master's degree	Student	1-3 years	Python	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
25970	494	50-54	Man	Sweden	Doctoral degree	Research Scientist	I have never written code	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
25971	277	45-49	Man	United States of America	Master's degree	Data Scientist	5-10 years	Python	NaN	SQL	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
25972	255	18-21	Man	India	Bachelor's degree	Business Analyst	I have never written code	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Removing Expectional Columns

```
raw_dataset_df.rename(columns={'Duration (in seconds)': 'Response Time',
                  'What is your age (# years)?': 'Age',
                  'What is your gender? - Selected Choice': 'Gender',
                  'In which country do you currently reside?': 'Country',
                  'What is the highest level of formal education that you have attained or plan to attain within the next 2 years?': 'Education',
                  'Select the title most similar to your current role (or most recent title if retired): - Selected Choice': 'Job Title',
                  'For how many years have you been writing code and/or programming?': 'Experience'}, inplace=True, errors='raise')
```

In [153]:
```
!pip install pandas_profiling
import pandas as pd
from pandas_profiling import ProfileReport
import seaborn as sns
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings('ignore')
%matplotlib inline
```

In [129]:
```
df = pd.read_csv("imp_dataset.csv")
df
```

Out[129]:

	Entity	Code	Year	5.b.1 - Proportion of individuals who own a mobile telephone, by sex (%) - IT_MOB_OWN - Male	5.b.1 - Proportion of individuals who own a mobile telephone, by sex (%) - IT_MOB_OWN - Both sexes	5.b.1 - Proportion of individuals who own a mobile telephone, by sex (%) - IT_MOB_OWN - Female
0	Afghanistan	AFG	2016	NaN	46.82	NaN
1	Algeria	DZA	2018	92.62	87.88	83.07
2	Armenia	ARM	2017	72.57	72.61	72.64
3	Armenia	ARM	2018	75.95	76.43	76.82
4	Azerbaijan	AZE	2014	77.21	73.92	70.66
...	...	...	...	...	...	...
261	Uruguay	URY	2018	75.81	76.69	77.53
262	Uzbekistan	UZB	2017	73.30	63.50	53.78
263	Uzbekistan	UZB	2018	77.40	68.10	58.85
264	Zambia	ZMB	2018	43.97	44.62	45.22
265	Zimbabwe	ZWE	2019	73.00	NaN	72.00

266 rows × 6 columns

In [131]:
```
df.sort_values("Year",ascending=True,ignore_index=True,)
df =df.drop(["Code",],axis=1)
df=df.fillna(0)
# df = df.rename(columns={"5.b.1 - Proportion of individuals who own a mobile telephone, by sex (%) - IT_MOB_OWN - B
df.columns=['Country','Year','Proportion with mobile(%) - male','Proportion with mobile(%) - both sexes','Proportion
```

In [132]:
```
df
df.to_csv("clean_data.csv")
```

In [133]:
```
yearwise=df.groupby("Year")
len(yearwise)
```

Out[133]: 7

In [154]:
```
df_2014=df.loc[df["Year"]==2014]
df_2015=df.loc[df["Year"]==2015]
df_2016=df.loc[df["Year"]==2016]
df_2017=df.loc[df["Year"]==2017]
df_2018=df.loc[df["Year"]==2018]
df_2019=df.loc[df["Year"]==2019]
```

In [155]:
```
df.drop([44],axis=0)
```

In [156]:
```
df_2015["Country"]
```

In [144]:
```
cont=list(df_2014["Country"])
for i,j in zip(df_2015["Country"],df_2015.index):
    if i not in cont:
        df_2015.drop(j,inplace=True)
#         print(i)
```

In [145]:
```
len(df_2015)
```

Out[145]: 22

In [146]:
```
# regression line accprding to year diff year
#piechart with contries both
```

In [ ]:
```
cont=list(df_2015["Country"])
for i,j in zip(df_2016["Country"],df_2016.index):
    if i not in cont:
        df_2016.drop(j,inplace=True)
len(df_2016)
# df_2016
```

In [ ]:
```

```

In [ ]:
```
cont=list(df_2014["Country"])
for i,j in zip(df_2017["Country"],df_2017.index):
    if i not in cont:
        df_2017.drop(j,inplace=True)
len(df_2017)
```

In [ ]:
```
cont=list(df_2014["Country"])
for i,j in zip(df_2018["Country"],df_2018.index):
    if i not in cont:
        df_2018.drop(j,inplace=True)
len(df_2018)
```

In [ ]:
```
# df_2014.to_csv("df_2014.csv")
```

In [ ]:
```
# df_2015.to_csv("df_2015.csv")
```

In [ ]:
```
# df_2016.to_csv("df_2016.csv")
```

In [ ]:
```
# df_2017.to_csv("df_2017.csv")
```

In [ ]:
```
# df_2018.to_csv("df_2018.csv")
```

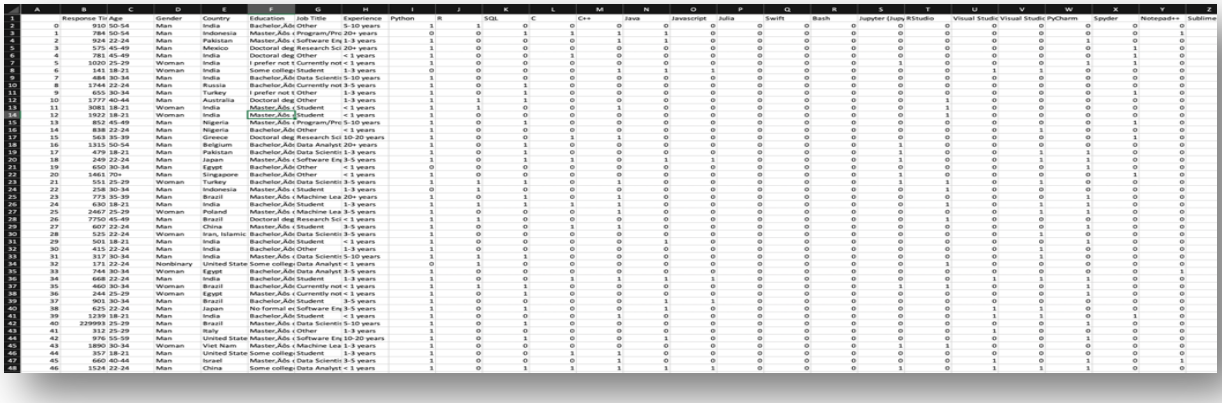
In [ ]:
```
df_2015 = df_2015.drop()
df_2 = pd.concat([df_2015,df_2016,df_2017,df_2018],ignore_index=True)
df_2.tail()
```

```
from sklearn import linear_model
reg = linear_model.LinearRegression()
reg.fit(nndf[['Year']],nndf['Mobile Access'])

num1 = np.array([i for i in range(2021, 2031)])
year = pd.DataFrame(num1)
predicted_vals = regC.predict(year)
predicted_vals

x = list(ndf['Year']) + list(num1)
y = list(ndf['Mobile Access']) + list(pr)

final_df = pd.DataFrame()
final_df['Year'] = x
final_df['Mobile Access'] = y
final_df

x = final_df['Year']
y = final_df['Mobile Access']
plt.figure(figsize=[10, 8])
plt.plot(x, y, 'o')

m, b = np.polyfit(x, y, 1)
plt.xlabel('\nYear')
plt.ylabel("Value (in lakh tones)")

plt.plot(x, m*x + b, color='red')
plt.grid(True)
plt.show()
```

Sample Data-

For Objective 1→

Raw Data



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1		Response Tir	Age	Gender	Country	Education	Job Title	Experience	Python	R	SQL	C	C++	Java	Javascript	Julia	Swift	Bash	Jupyter (Jupy	RStudio	Visual Studio	Visual Studio	PyCharm	Spyder	Notepad++	Sublime
2	0	910	50-54	Man	India	Bachelor,Ado	Other	5-10 years	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	1	784	50-54	Man	Indonesia	Master,Ado	Program/Pro	20+ years	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	1
4	2	924	22-24	Man	Pakistan	Master,Ado	Software Eng	1-3 years	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	0	0
5	3	575	45-49	Man	Mexico	Doctoral deg	Research Sci	20+ years	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
6	4	781	45-49	Man	India	Doctoral deg	Other	< 1 years	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0
7	5	1020	25-29	Woman	India	I prefer not t	Currently not	< 1 years	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	0
8	6	141	18-21	Woman	India	Some college	Student	1-3 years	0	0	0	0	1	1	1	0	0	0	0	0	0	1	1	0	0	0
9	7	484	30-34	Man	India	Bachelor,Ado	Data Scientis	5-10 years	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	8	1744	22-24	Man	Russia	Bachelor,Ado	Currently not	3-5 years	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	9	655	30-34	Man	Turkey	I prefer not t	Other	1-3 years	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
12	10	1777	40-44	Man	Australia	Doctoral deg	Other	1-3 years	1	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
13	11	3081	18-21	Woman	India	Master,Ado	Student	< 1 years	1	1	0	0	1	1	0	0	0	0	0	0	1	0	0	0	0	0
14	12	1922	18-21	Woman	India	Master,Ado	Student	< 1 years	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
15	13	852	45-49	Man	Nigeria	Master,Ado	Program/Pro	5-10 years	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
16	14	838	22-24	Man	Nigeria	Bachelor,Ado	Other	< 1 years	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
17	15	563	35-39	Man	Greece	Doctoral deg	Research Sci	10-20 years	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	1	0
18	16	1315	50-54	Man	Belgium	Bachelor,Ado	Data Analyst	20+ years	1	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
19	17	479	18-21	Man	Pakistan	Bachelor,Ado	Data Scientis	1-3 years	1	0	1	0	0	0	0	0	0	0	0	1	0	0	1	1	0	0
20	18	249	22-24	Man	Japan	Master,Ado	Software Eng	3-5 years	1	0	1	1	0	0	1	1	0	0	0	1	0	0	1	1	0	0
21	19	650	30-34	Man	Egypt	Bachelor,Ado	Other	< 1 years	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
22	20	1461	70+	Man	Singapore	Bachelor,Ado	Other	< 1 years	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0
23	21	551	25-29	Woman	Turkey	Bachelor,Ado	Data Scientis	3-5 years	1	1	1	0	1	0	0	0	0	0	1	1	0	1	0	0	0	0
24	22	258	30-34	Man	Indonesia	Master,Ado	Student	1-3 years	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
25	23	773	35-39	Man	Brazil	Master,Ado	Machine Lea	20+ years	1	0	1	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0
26	24	630	18-21	Man	India	Bachelor,Ado	Student	1-3 years	1	1	1	1	1	0	0	0	0	0	0	1	0	1	1	0	0	0
27	25	2467	25-29	Woman	Poland	Master,Ado	Machine Lea	3-5 years	1	0	0	0	1	0	0	0	0	0	0	0	0	1	1	0	0	0
28	26	7750	45-49	Man	Brazil	Doctoral deg	Research Sci	< 1 years	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
29	27	607	22-24	Man	China	Master,Ado	Student	3-5 years	1	0	0	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0
30	28	525	22-24	Woman	Iran, Islamic	Bachelor,Ado	Data Scientis	3-5 years	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
31	29	501	18-21	Man	India	Bachelor,Ado	Student	< 1 years	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0
32	30	415	22-24	Man	India	Bachelor,Ado	Other	1-3 years	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
33	31	317	30-34	Man	India	Master,Ado	Data Scientis	5-10 years	1	1	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
34	32	171	22-24	Nonbinary	United State	Some college	Data Analyst	< 1 years	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
35	33	744	30-34	Woman	Egypt	Bachelor,Ado	Data Analyst	3-5 years	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
36	34	668	22-24	Man	India	Bachelor,Ado	Student	1-3 years	1	0	0	1	1	1	1	0	0	0	0	0	1	1	1	0	0	0
37	35	460	30-34	Woman	Brazil	Bachelor,Ado	Currently not	< 1 years	1	1	1	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0
38	36	244	25-29	Woman	Egypt	Master,Ado	Currently not	< 1 years	1	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
39	37	901	30-34	Man	Brazil	Bachelor,Ado	Student	3-5 years	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
40	38	625	22-24	Man	Japan	No formal ed	Software Eng	3-5 years	1	0	1	0	0	1	0	0	0	0	0	0	1	1	0	0	0	0
41	39	1239	18-21	Man	India	Bachelor,Ado	Student	< 1 years	1	0	0	0	0	1	0	0	0	0	0	0	1	1	0	0	1	0
42	40	229993	25-29	Man	Brazil	Master,Ado	Data Scientis	5-10 years	1	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
43	41	312	25-29	Man	Italy	Master,Ado	Other	1-3 years	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
44	42	976	55-59	Man	United State	Master,Ado	Software Eng	10-20 years	1	0	1	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0
45	43	1890	30-34	Woman	Viet Nam	Master,Ado	Machine Lea	1-3 years	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
46	44	357	18-21	Man	United State	Some college	Student	1-3 years	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
47	45	660	40-44	Man	Israel	Master,Ado	Data Scientis	3-5 years	1	0	0	0	1	0	0	0	0	0	0	0	1	0	0	1	0	1
48	46	1524	22-24	Man	China	Some college	Data Analyst	< 1 years	1	0	1	1	1	1	1	0	0	0	1	0	1	1	1	0	0	0

Tools used by males and females.

Tools	Men	Women
Python	17582	3877
R	4072	1145
SQL	8449	2101
C	3633	991
C++	4465	972
Java	3691	975
Javascript	3563	677
Julia	268	24
Swift	204	26
Bash	1909	220
Jupyter (Jupyt	4527	847
RStudio	3627	1053
Visual Studio	3284	756
Visual Studio	8379	1476
PyCharm	6071	1238
Spyder	3038	686
Notepad++	3077	784
Sublime Text	2279	485
Vim / Emacs	1439	147

For Objective 2→

Data for male and female job with Age

Response Tir	Age	Gender	Country	Education	Job Title	Int Age
910	50-54	Man	India	Bachelor,Äôs degree	Other	50
784	50-54	Man	Indonesia	Master,Äôs degree	Program/Project Manager	50
924	22-24	Man	Pakistan	Master,Äôs degree	Software Engineer	22
575	45-49	Man	Mexico	Doctoral degree	Research Scientist	45
781	45-49	Man	India	Doctoral degree	Other	45
1020	25-29	Woman	India	I prefer not to answer	Currently not employed	25
141	18-21	Woman	India	Some college/university study without earning a bachelor,Äôs degree	Student	18
484	30-34	Man	India	Bachelor,Äôs degree	Data Scientist	30
1744	22-24	Man	Russia	Bachelor,Äôs degree	Currently not employed	22
655	30-34	Man	Turkey	I prefer not to answer	Other	30
1777	40-44	Man	Australia	Doctoral degree	Other	40
3081	18-21	Woman	India	Master,Äôs degree	Student	18
1922	18-21	Woman	India	Master,Äôs degree	Student	18
852	45-49	Man	Nigeria	Master,Äôs degree	Program/Project Manager	45
838	22-24	Man	Nigeria	Bachelor,Äôs degree	Other	22
563	35-39	Man	Greece	Doctoral degree	Research Scientist	35
1315	50-54	Man	Belgium	Bachelor,Äôs degree	Data Analyst	50
479	18-21	Man	Pakistan	Bachelor,Äôs degree	Data Scientist	18
249	22-24	Man	Japan	Master,Äôs degree	Software Engineer	22
650	30-34	Man	Egypt	Bachelor,Äôs degree	Other	30
1461	70+	Man	Singapore	Bachelor,Äôs degree	Other	70
551	25-29	Woman	Turkey	Bachelor,Äôs degree	Data Scientist	25
258	30-34	Man	Indonesia	Master,Äôs degree	Student	30
773	35-39	Man	Brazil	Master,Äôs degree	Machine Learning Engineer	35
630	18-21	Man	India	Bachelor,Äôs degree	Student	18
2467	25-29	Woman	Poland	Master,Äôs degree	Machine Learning Engineer	25
7750	45-49	Man	Brazil	Doctoral degree	Research Scientist	45
607	22-24	Man	China	Master,Äôs degree	Student	22
525	22-24	Woman	Iran, Islamic	Bachelor,Äôs degree	Data Scientist	22
501	18-21	Man	India	Bachelor,Äôs degree	Student	18
415	22-24	Man	India	Bachelor,Äôs degree	Other	22
317	30-34	Man	India	Master,Äôs degree	Data Scientist	30
171	22-24	Nonbinary	United State	Some college/university study without earning a bachelor,Äôs degree	Data Analyst	22
744	30-34	Woman	Egypt	Bachelor,Äôs degree	Data Analyst	30
668	22-24	Man	India	Bachelor,Äôs degree	Student	22
460	30-34	Woman	Brazil	Bachelor,Äôs degree	Currently not employed	30
244	25-29	Woman	Egypt	Master,Äôs degree	Currently not employed	25
901	30-34	Man	Brazil	Bachelor,Äôs degree	Student	30
625	22-24	Man	Japan	No formal education past high school	Software Engineer	22
1239	18-21	Man	India	Bachelor,Äôs degree	Student	18
229993	25-29	Man	Brazil	Master,Äôs degree	Data Scientist	25
312	25-29	Man	Italy	Master,Äôs degree	Other	25
976	55-59	Man	United State	Master,Äôs degree	Software Engineer	55
1890	30-34	Woman	Viet Nam	Master,Äôs degree	Machine Learning Engineer	30
357	18-21	Man	United State	Some college/university study without earning a bachelor,Äôs degree	Student	18
660	40-44	Man	Israel	Master,Äôs degree	Data Scientist	40
1524	22-24	Man	China	Some college/university study without earning a bachelor,Äôs degree	Data Analyst	22
453	18-21	Woman	Egypt	Bachelor,Äôs degree	Student	18

PERCENTAGE DATA OF INDIA OF BOTH GENDERS IN OCCUPATION IN SAME FIELD IN %.

1	Occupation	Men	Women
2	Student	8.121469	2.89548
3	Data Scientist	2.87194	0.553202
4	Software Engineer	2.248117	0.596359
5	Other	1.502668	0.580665
6	Data Analyst	1.647834	0.537508
7	Currently not employed	1.871469	0.478657
8	Machine Learning Engineer	1.251569	0.219711
9	Research Scientist	0.635593	0.133396
10	Business Analyst	0.455116	0.066698
11	Program/Project Manager	0.192247	0.094162
12	Data Engineer	0.48258	0.219711
13	Product Manager	0.192247	0.01177
14	Statistician	0.094162	0.027464
15	DBA/Database Engineer	0.105932	0.058851
16	Developer Relations/Advocacy	0.047081	0.02354

PERCENTAGE DATA OF WORLD OF BOTH GENDERS IN OCCUPATION IN SAME FIELD IN %.

	A	B	C
1	Occupation	Men	Women
2	Student	19.981952	6.175455
3	Data Scientist	11.656466	2.291274
4	Software Engineer	8.023384	1.43597
5	Other	7.556497	1.683145
6	Data Analyst	6.901287	1.981325
7	Currently not employed	5.802731	1.816541
8	Machine Learning Engineer	5.041588	0.702291
9	Research Scientist	4.837571	1.075016
10	Business Analyst	3.05634	0.686598
11	Program/Project Manager	2.911174	0.357031
12	Data Engineer	2.201036	0.384495
13	Product Manager	1.043628	0.176554
14	Statistician	0.937696	0.278562
15	DBA/Database Engineer	0.553202	0.078468
16	Developer Relations/Advocacy	0.30995	0.062775

DATA OF COMPANY EMPLOYEES IN %

Year	Male	Female
2010	80	20
2011	79	21
2012	79	21
2013	78	22
2014	78	22
2015	76	24
2016	75	25
2017	74	26
2018	71	29
2019	70	30
2020	67	31
2021	64	32

For Objective 3→

Data for investment and compensation of different ages.

Response Tir	Age	Gender	Country	Education	Job Title	Experience	Investment	Compensatic	Int Age
910	50-54	Man	India	Bachelor,Äôs	Other	5-10 years	$100-$999	25,000-29,99	50
784	50-54	Man	Indonesia	Master,Äôs d	Program/Pro	20+ years	$0 ($USD)	60,000-69,99	50
924	22-24	Man	Pakistan	Master,Äôs d	Software Eng	1-3 years	$0 ($USD)	$0-999	22
575	45-49	Man	Mexico	Doctoral deg	Research Sci	20+ years	$0 ($USD)	30,000-39,99	45
781	45-49	Man	India	Doctoral deg	Other	< 1 years	$1000-$9,99	30,000-39,99	45
1020	25-29	Woman	India	I prefer not t	Currently not	< 1 years			25
141	18-21	Woman	India	Some college	Student	1-3 years			18
484	30-34	Man	India	Bachelor,Äôs	Data Scientis	5-10 years	$1-$99	15,000-19,99	30
1744	22-24	Man	Russia	Bachelor,Äôs	Currently not	3-5 years			22
655	30-34	Man	Turkey	I prefer not t	Other	1-3 years	$0 ($USD)	$0-999	30
1777	40-44	Man	Australia	Doctoral deg	Other	1-3 years	$1-$99	70,000-79,99	40
3081	18-21	Woman	India	Master,Äôs d	Student	< 1 years			18
1922	18-21	Woman	India	Master,Äôs d	Student	< 1 years			18
852	45-49	Man	Nigeria	Master,Äôs d	Program/Pro	5-10 years	$100-$999	2,000-2,999	45
838	22-24	Man	Nigeria	Bachelor,Äôs	Other	< 1 years	$0 ($USD)	$0-999	22
563	35-39	Man	Greece	Doctoral deg	Research Sci	10-20 years	$0 ($USD)	10,000-14,99	35
1315	50-54	Man	Belgium	Bachelor,Äôs	Data Analyst	20+ years	$0 ($USD)	2,000-2,999	50
479	18-21	Man	Pakistan	Bachelor,Äôs	Data Scientis	1-3 years	$0 ($USD)	$0-999	18
249	22-24	Man	Japan	Master,Äôs d	Software Eng	3-5 years			22
650	30-34	Man	Egypt	Bachelor,Äôs	Other	< 1 years	$0 ($USD)	5,000-7,499	30
1461	70+	Man	Singapore	Bachelor,Äôs	Other	< 1 years	$1-$99	20,000-24,99	70
551	25-29	Woman	Turkey	Bachelor,Äôs	Data Scientis	3-5 years	$100-$999	1,000-1,999	25
258	30-34	Man	Indonesia	Master,Äôs d	Student	1-3 years			30
773	35-39	Man	Brazil	Master,Äôs d	Machine Lea	20+ years	$0 ($USD)	100,000-124,	35
630	18-21	Man	India	Bachelor,Äôs	Student	1-3 years			18
2467	25-29	Woman	Poland	Master,Äôs d	Machine Lea	3-5 years	$0 ($USD)	25,000-29,99	25
7750	45-49	Man	Brazil	Doctoral deg	Research Sci	< 1 years	$0 ($USD)	$0-999	45
607	22-24	Man	China	Master,Äôs d	Student	3-5 years			22
525	22-24	Woman	Iran, Islamic	Bachelor,Äôs	Data Scientis	3-5 years	$0 ($USD)	$0-999	22
501	18-21	Man	India	Bachelor,Äôs	Student	< 1 years			18
415	22-24	Man	India	Bachelor,Äôs	Other	1-3 years	$100-$999	7,500-9,999	22
317	30-34	Man	India	Master,Äôs d	Data Scientis	5-10 years	$100,000 or	100,000-124,	30
171	22-24	Nonbinary	United State	Some college	Data Analyst	< 1 years			22
744	30-34	Woman	Egypt	Bachelor,Äôs	Data Analyst	3-5 years	$0 ($USD)	7,500-9,999	30
668	22-24	Man	India	Bachelor,Äôs	Student	1-3 years			22
460	30-34	Woman	Brazil	Bachelor,Äôs	Currently not	< 1 years			30
244	25-29	Woman	Egypt	Master,Äôs d	Currently not	< 1 years			25
901	30-34	Man	Brazil	Bachelor,Äôs	Student	3-5 years			30
625	22-24	Man	Japan	No formal ed	Software Eng	3-5 years	$0 ($USD)	2,000-2,999	22
1239	18-21	Man	India	Bachelor,Äôs	Student	< 1 years			18
229993	25-29	Man	Brazil	Master,Äôs d	Data Scientis	5-10 years			25
312	25-29	Man	Italy	Master,Äôs d	Other	1-3 years	$1-$99	30,000-39,99	25
976	55-59	Man	United State	Master,Äôs d	Software Eng	10-20 years	$1000-$9,99	15,000-19,99	55
1890	30-34	Woman	Viet Nam	Master,Äôs d	Machine Lea	1-3 years	$1000-$9,99	4,000-4,999	30
357	18-21	Man	United State	Some college	Student	1-3 years			18
660	40-44	Man	Israel	Master,Äôs d	Data Scientis	3-5 years	$10,000-$99	100,000-124,	40
1524	22-24	Man	China	Some college	Data Analyst	< 1 years	$0 ($USD)	$0-999	22

Data of Compensation Taken By Both Gender in same Field.

	A	B	C
1	Occupation	Men	Women
2	Student	0	0
3	Data Scientis	30120.94	31139.59
4	Software Eng	57561.24	37832.9
5	Other	42907.15	28023.65
6	Data Analyst	38472.35	21728.81
7	Currently not	87405.06	69256.1
8	Machine Lear	51752.98	31229.41
9	Research Scie	63836.01	46576.47
10	Business Anal	0	0
11	Program/Proj	48922.7	31279.59
12	Data Enginee	61955.22	49600
13	Product Man	48204.97	30646.86
14	Statistician	41660.14	37821.66
15	DBA/Databas	51672	31184.21
16	Developer Re	57747.76	40051.49

Data of Investment Done By Both Gender in %.

Job Title	Male	Female
Business Ana	4.1950196	0.9283603
Currently no	0	0
DBA/Databas	0.570429	0.060247
Data Analyst	7.2160325	2.2196513
Data Enginee	4.1855323	0.5700882
Data Scientis	28.356561	3.4940696
Developer R	0.2999567	0.1164889
Machine Lea	9.6982026	0.5584989
Other	8.40819	0.9300078
Product Mar	2.4191963	0.3819051
Program/Pro	5.4029703	0.511091
Research Sci	7.3484849	0.8808388
Software Eng	9.5721413	0.6571779
Statistician	0.9141578	0.1047008
Student	0	0

For Objective 4→

Proportion of individuals who own a mobile telephone, by sex (%)

	Country	Year	Proportion w	Proportion with r
0	Afghanistan	2016	0	0
1	Algeria	2018	92.62	83.07
2	Armenia	2017	72.57	72.64
3	Armenia	2018	75.95	76.82
4	Azerbaijan	2014	77.21	70.66
5	Azerbaijan	2015	77.81	70.47
6	Azerbaijan	2016	77.01	69.84
7	Azerbaijan	2017	88.38	79.72
8	Azerbaijan	2018	87.75	79.34
9	Azerbaijan	2019	87.64	79.43
10	Bahrain	2015	100	100
11	Bahrain	2016	100	100
12	Bahrain	2017	100	100
13	Bahrain	2018	100	100
14	Bahrain	2019	100	100
15	Bangladesh	2017	54.3	30.91
16	Bangladesh	2019	0	0
17	Belarus	2017	91.48	94.99
18	Belarus	2018	93.69	95.7
19	Belarus	2019	93.95	96.26

Data for Cellular data user from year 1961 to 2020 in millions.

	Year	Mobile Access
0	1961	0.000000
1	1962	0.000000
2	1963	0.000000
3	1964	0.000000
4	1965	0.000000
...	...	...
65	2026	857.205303
66	2027	880.897864
67	2028	904.590425
68	2029	928.282985
69	2030	951.975546

70 rows × 2 columns

MATHEMATICAL MODEL

STATISTICAL APPROACH:

Statistic can be characterized as a part of math that changes information into helpful data for the chiefs.

There are certain types of method in statistic:

- Variance:

The term change alludes to a factual estimation of the spread between numbers in an informational collection. All the more explicitly, change estimates how far each number in the set is from the mean and subsequently from each and every other number in the set. Change is regularly portrayed by this image: σ2. It is utilized

by the two examiners and brokers to decide unpredictability and market security.

- Median:

Middle addresses the centre incentive for any gathering. It is the place where a large portion of the information is more and a large portion of the information is less. Middle assists with addressing an enormous number of informative elements with a solitary element. The middle is the most straightforward factual measure to compute. For estimation of middle, the information must be organized in climbing request, and afterward the middlemost information point addresses the middle of the information.

- Mean:

Mean is a measurable idea that conveys a significant importance in finance and is utilized in different monetary fields and business valuation.

$$\text{mean} = \frac{\sum_{i=1}^{N} X_i}{N}$$

- Standard Deviation:

Standard deviation is the level of scattering or the disperse of the information directs relative toward its mean, in illustrative measurements. It tells how the qualities are spread across the information test and it is the proportion of the variety of the items from the mean.

$$\mathrm{SD} = \sqrt{\frac{\sum |x - \mu|^2}{N}}$$

Kurtosis- The sharpness of the pinnacle of a recurrence circulation bend.

Skewness- Skewness is a proportion of the imbalance of the likelihood dispersion of a genuine esteemed irregular variable with regards to its mean. The skewness worth can be positive, zero, negative, or vague.

Range- The distinction between the least and most noteworthy qualities

Minimum- Least means the most reduced sum or level of something. Least can be a thing or a descriptive word and it has a few more explicit implications, which are all related somehow or another to its essential

importance. At the point when least is utilized as a thing, its plural can be essentials or, less regularly, minima.

Maximum- The most extreme worth of a capacity is where a capacity arrives at its most noteworthy point, or vertex, on a chart. ... For all intents and purposes, observing the most extreme worth of a capacity can be utilized to decide greatest benefit or most extreme region.

INFERENTIAL STATISTICS:

Includes utilizing test information to make a surmising or reach an inference of the populace. It utilizes likelihood to decide how certain we can be that the ends we make are right.

Hypothesis Testing:

It is the strategy utilized in settling on measurable choices utilizing exploratory information. It is fundamentally a suspicion that we by and large make about populace boundaries. We use Hypothesis testing at whatever point we are attempting to draw a few derivations on the total dataset by thinking about just an example of the entire populace. Invalid theory and substitute speculation are the two principal credits for speculation testing.

Test of Hypothesis concerning single mean

(Single Population)

Step 1: Null Hypothesis H_0: $\mu = \mu_0$

Alternative Hypothesis: $\mu > \mu_0$ (Right tailed);

$\mu < \mu 0$ (Left tailed);

$\mu =/ \mu 0$ (Two tailed)

Level of Significance: α

Sample Size: n

Step 2: Test Statistic

Case 1: If population variance is known:

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

Case 2: If population variance in unknown and n < 30:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}, \text{ where } s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n-1}$$

Step 3: Criteria of Rejection: Reject the null hypothesis if

Case 1: $z > z\alpha$ (Right Tailed); $z < -z\alpha$ (Left Tailed);

$|z| > z\alpha/2$ (Two Tailed)

Case2:

$t > t\alpha,n-1$ (Right Tailed); $t < -t\alpha,n-1$ (Left Tailed);

$|t| > t\alpha/2,n-1$ (Two Tailed)

Step 4: Calculation

Step5: Decision

Test of Hypothesis concerning difference of means (Two Populations):

Step 1: Null Hypothesis H_0: $\mu_1 - \mu_2 = d$

Alternative Hypothesis: $\mu_1 - \mu_2 > d$ (Right tailed);

$\mu_1 - \mu_2 < d$ (Left tailed);

$\mu_1 - \mu_2 \neq d$ (Two tailed) Level of Significance: α

Sample Size: Sample 1 - n_1 and Sample 2 – n_2

Step 2: Test Statistic

Case 1: If population variance is known:

$$z = \frac{\overline{x_1} - \overline{x_2} - \delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Where, and are population variances of sample 1 and sample 2 respectively.

Case 2: If population variance in unknown and n < 30:

$$t = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Where;

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2},$$

s_1^2 and s_2^2 are the sample variances of sample 1 and 2 respectively

Here, $s_p^2 = \frac{s_1^2 + s_2^2}{2}$ if $n_1 = n_2$

Here, if $n_1 = n_2$

Step 3: Criteria of Rejection

Reject the null hypothesis if

Case 1: $z > z_{\alpha}$ (Right Tailed);

$z < -z_{\alpha}$ (Left Tailed);

$| z | > z_{\alpha/2}$ (Two Tailed)

Case 2: (Right Tailed);

Step 4: Calculation

Step5: Decision (Left Tailed); (Two Tailed)

Test of Hypothesis concerning Proportion

(Single Population)

Step 1: Null Hypothesis H_0: $p = p_0$

Alternative Hypothesis: $p > p_0$ (Right tailed); $p < p_0$ (Left tailed); $p \neq p_0$

(Two tailed)

Level of Significance: α

Sample Size: n

Step 2: Test Statistic:

$$z = \frac{x - np_0}{\sqrt{np_0(1-p_0)}}$$

Step 3: Criteria of Rejection

Reject the null hypothesis if $z > z_\alpha$ (Right Tailed); $z < -z_\alpha$ (Left Tailed); $|z| > z_{\alpha/2}$ (Two Tailed)

Step 4: Calculation

Step5: Decision

Test of Hypothesis concerning Two Variances

Step 1: Null Hypothesis H_0: $s_1^2 = s_2^2$

Alternative Hypothesis: $s_1^2 > s_2^2$ (Right tailed); $s_1^2 < s_2^2$ (Left tailed); $s_1^2 \neq s_2^2$ (Two tailed)

Level of Significance: α

Sample Size: Sample 1 - n_1 and Sample 2 – n_2

Step 2: Test Statistic:

$$F = \frac{S_i^2}{S_j^2}, \text{ where } I > j$$

Step 3: Criteria of Rejection

Reject the null hypothesis if $F > F_{\alpha,n_1-1,n_2-1}$ (Right Tailed); $F > F_{\alpha,n_2-1,n_1-1}$ (Left Tailed); $F > F_{\frac{\alpha}{2},n_i-1,n_j-1}$ (Two Tailed)

Step 4: Calculation Step5: Decision

Note: The numerator in F is always greater than denominator and the in the sequence of degree of freedom also, the larger sample size comes first.

Analysis of Variance (ANOVA)

One Way Classification

Step 1: Null Hypothesis H_0: Population means are all equal i.e., $\mu_1 = \mu_2 = ... = \mu_n$ ($\alpha_i = 0$, for all i)

Alternative Hypothesis: All population means are not equal i.e.,

$\mu_1 \neq \mu_2 \neq ... \neq \mu_n$

($\alpha_i = 0$, for at least i)

Level of Significance: α

Step 2: Test Statistic:

$$F = \frac{MS\ (Tr)}{MSE}$$

Treatment Mean Squaure $MS\ (Tr) = \frac{SS\ (Tr)}{k-1}$

Error Mean Squaure $MSE = \frac{SSE}{k\ (n-1)}$

Total sum of squares $SST = \sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}^2 - \frac{1}{N}T_{..}^2$

Treatment sum of squares $SS(Tr) = \frac{1}{n}\sum_{i=1}^{k} T_{i\cdot}^2 - \frac{1}{N}T_{..}^2$

Error sum of squares $SSE = SST - SS(Tr)$

Where $T_{i\cdot}$ is the total of i[th] row and $T_{..}$ is the grand total

Step 3: Criteria of Rejection: Reject the null hypothesis if

$$F > F_{\alpha,k-1,N-k}$$

Step 4: Calculation

Source of Variation	Degrees of freedom	Sum of Squares (SS)	Mean Square (MS)	F (Test Statistic)
Treatments	k - 1	SS (Tr)	MS (Tr)	$\frac{MS\ (Tr)}{MSE}$
Errors	N - k	SSE	MSE	
Total	N - 1	SST		

Step5: Decision

Inferences on a Population Correlation Coefficient

Step 1: Null Hypothesis H_0: r = 0

Alternative Hypothesis: r > 0 (Right tailed); r < 0 (Left tailed); r ≠ 0

(Two tailed)

Level of Significance: α

Step 2: Test Statistic:

$$\Delta_r = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \text{ (For small samples)}$$

General Method:

$$Z = \frac{\sqrt{n-3}}{2} \cdot \ln\frac{1+r}{1-r}$$

Where

$$r = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}}$$

Step 3: Criteria of Rejection

Small Samples: Reject the null hypothesis if $\Delta_r > t_{\alpha,n-2}$ (Right Tailed); $\Delta_r < -t_{\alpha,n-2}$ (Left Tailed); $\Delta_r > t_{\alpha/2,n-2}$ or $\Delta_r < -t_{\alpha/2,n-2}$ (Two Tailed)

General Method using Z statistic: By usual method

Step 4: Calculation

Step5: Decision

Inferences on a Population Regression Line

Step 1: Null Hypothesis H_0: $b = b_0$

Alternative Hypothesis: $b > b_0$ (Right tailed); $b < b_0$ (Left tailed); $b \neq b_0$ (Two tailed)

Level of Significance: α

Step 2: Test Statistic: $t = \frac{\hat{\beta}-\beta_0}{\hat{\sigma}}\sqrt{\frac{(n-2)S_{xx}}{n}}$

Where $\hat{\beta} = \frac{S_{xy}}{S_{xx}}$ and $\hat{\sigma} = \sqrt{\frac{1}{n}\left(S_{yy} - \hat{\beta}\cdot S_{xy}\right)}$

Step 3: Criteria of Rejection

Small Samples: Reject the null hypothesis if $t > t_{\alpha,n-2}$ (Right Tailed); $t < -t_{\alpha,n-2}$ (Left Tailed); $t > t_{\alpha/2,n-2}$ or $t < -t_{\alpha/2,n-2}$ (Two Tailed)

Step 4: Calculation

Step 5: Decision

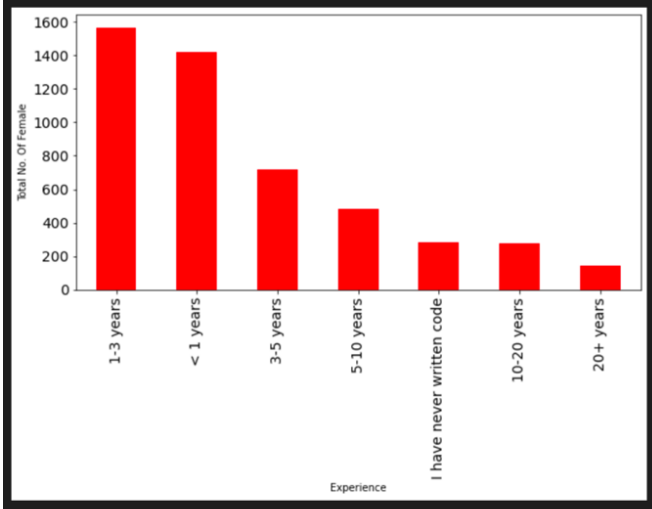
Steps and Calculation with Graphs:

PS 1.1 According to our Problem Statement, we are going to use **Correlation** and our problem statement is "**To analyse and study Information technology field expertise in Machine learning and Data science.**"

Our Statement- "Our statement is to see correlation between Employees age with total no. of expertise."

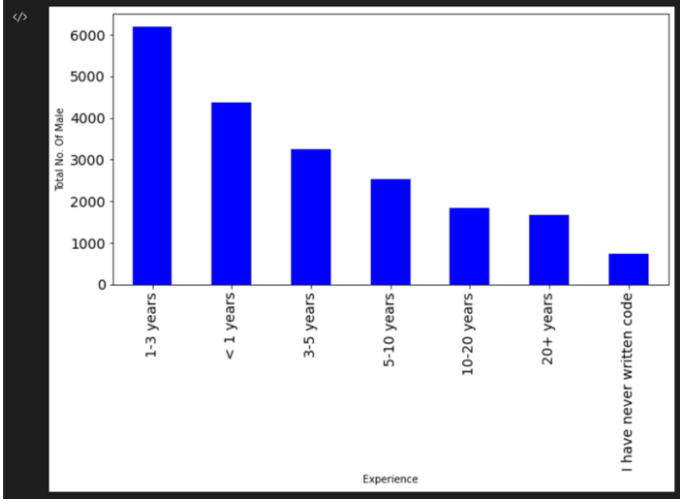
```
gender_exper_series["Woman"].plot.bar(color="red", figsize=(10, 5), fontsize=14, ylabel="Total No. Of Female")
```

```
<AxesSubplot:xlabel='Experience', ylabel='Total No. Of Female'>
```




```
gender_exper_series = dataset.groupby("Gender")['Experience'].value_counts()
gender_exper_series["Man"].plot.bar(color="blue", figsize=(10, 5), fontsize=14, ylabel="Total No. Of Male")
```

```
<AxesSubplot:xlabel='Experience', ylabel='Total No. Of Male'>
```



```
dataset.loc[:, ["Age", "Total Skill Set"]].corr()
```

	Age	Total Skill Set
Age	1.00000	0.08176
Total Skill Set	0.08176	1.00000

àSo the correlation between age and skill set is 0.8999.

“So it is not necessary that age matters in knowledge it means small child has more knowledge than 30 year old man on that field.”

PS 1.2 According to our Problem Statement, we are going to use **Test of Hypothesis concerning difference of means (Two Populations)** and our prediction is “**In Men and Women, more languages are learned by men comparatively women.**”

So here are some calculation:

$$\mu 1 = Men$$

$$\mu 2 = Women$$

Step 1: Null Hypothesis H_0: $\mu_1 - \mu_2 = d$

Alternative Hypothesis: $\mu_1 - \mu_2 > d$ (Right tailed);

$\mu_1 - \mu_2 < d$ (Left tailed);

$\mu_1 - \mu_2 \neq d$ (Two tailed) Level of Significance: α

Sample Size: Sample 1 - n_1 and Sample 2 – n_2

$$H0=\mu 1 - \mu 2 = 0$$

$$Ha= \mu 1 - \mu 2 \neq 0$$

Level of Significance: $\boldsymbol{\alpha = 0.05}$

N1=162 and N2=162 (Two Tailed)

Step 2: Test Statistic:

N1+N2=324>30 large sample.

$\sigma 1^2 \; and \; \sigma 2^2$ are not known.

$$z = \frac{\overline{x_1} - \overline{x_2} - \delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Step 4: Calculation

$$|z| > z\alpha/2$$

$$|z| = z(0.05) = 1.960$$

X1 bar =2053.64

X2 bar =390.19

$$\sigma 1^2 = 88131.92$$

$$\sigma 2^2 = 404148.45$$

$$z = \frac{2053 - 390}{\sqrt{\frac{(88131)^2 - (404148)^2}{162}}}$$

$$z = \frac{1663}{56806}$$

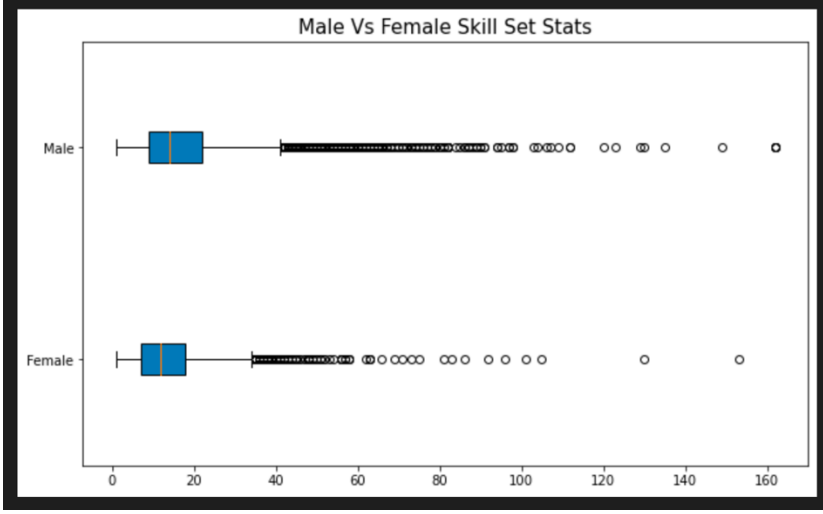
Z=0.029

Step5: Decision

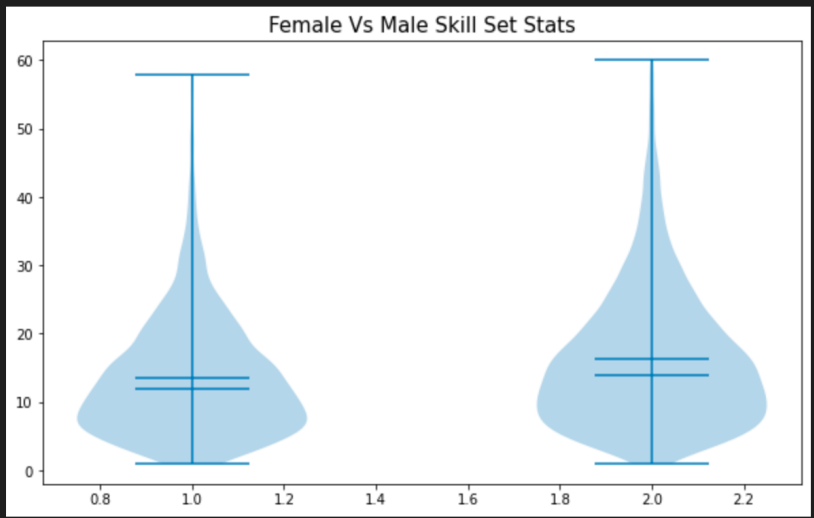
$Z\alpha\backslash 2 > |z|$

So, "In Men and Women, more languages are learned by men comparatively women."

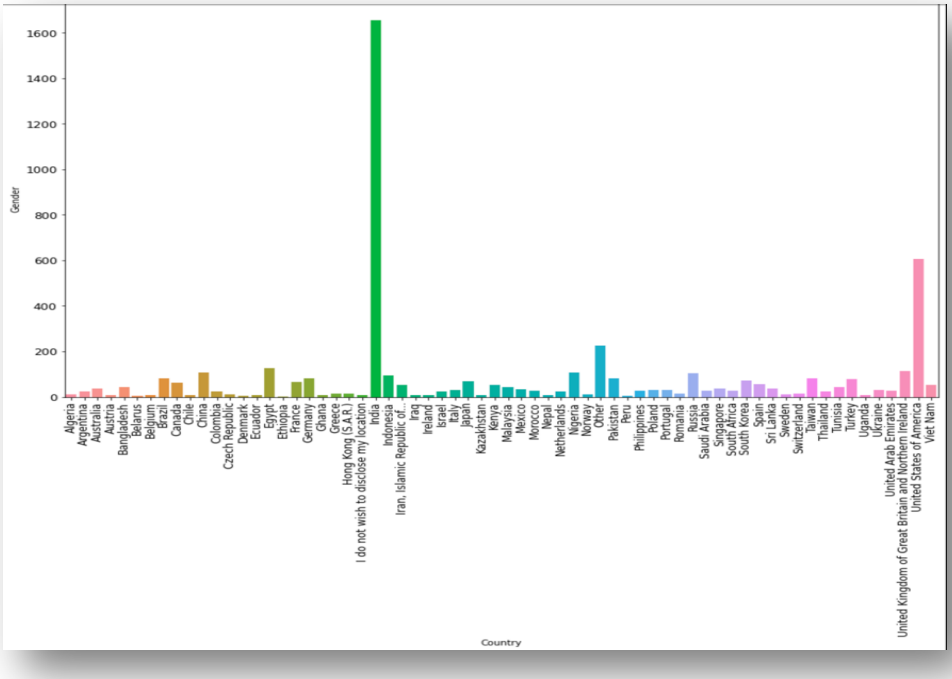
```
women_skillset_series = dataset.loc[(dataset["Gender"] == "Woman") & (dataset["Total Skill Set"] != 0)]["Total Skill Set"]
men_skillset_series = dataset.loc[(dataset["Gender"] == "Man") & (dataset["Total Skill Set"] != 0)]["Total Skill Set"]
plt.figure(figsize=(10,6))
plt.boxplot([women_skillset_series, men_skillset_series], labels=["Female","Male"], patch_artist = True, vert = 0)
plt.title("Male Vs Female Skill Set Stats", size=15)
plt.show()
```



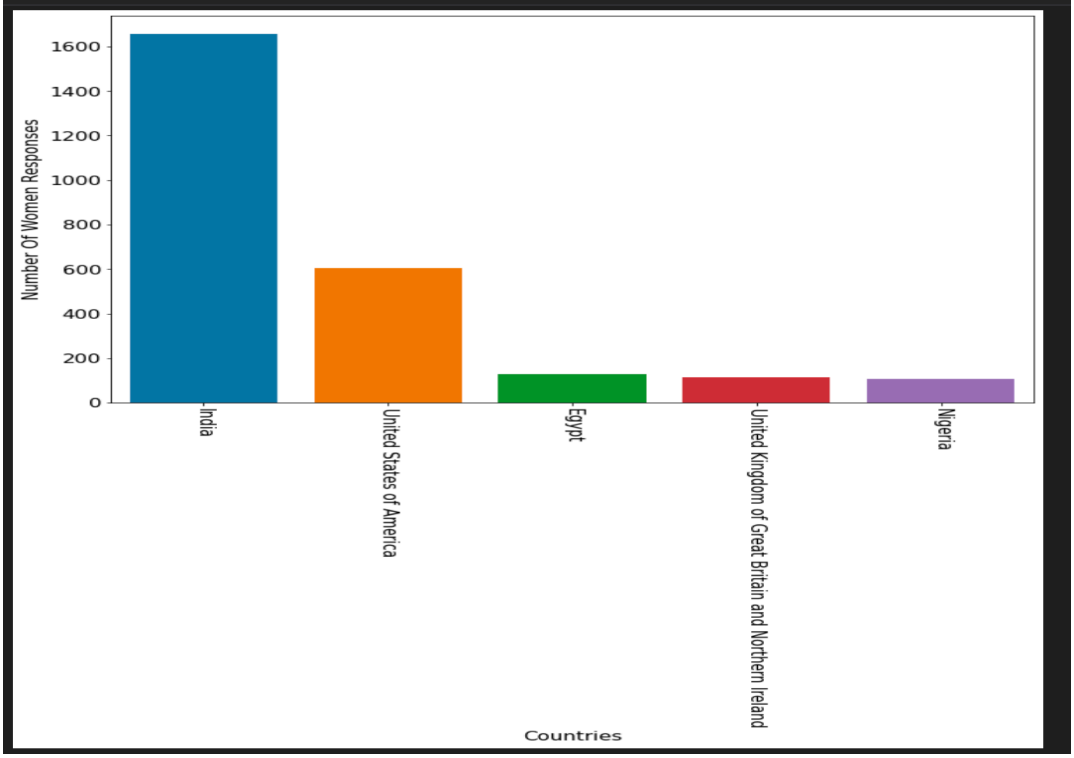

```
zoomed_women_skillset_series = women_skillset_series.loc[women_skillset_series <= 60]
zoomed_men_skillset_series = men_skillset_series.loc[men_skillset_series <= 60]
plt.figure(figsize=(10,6))
plt.violinplot([zoomed_women_skillset_series, zoomed_men_skillset_series],
               showmeans=True, showextrema=True, showmedians=True,
               points=100)
plt.title("Female Vs Male Skill Set Stats", size=15)
plt.show()
```



PS 1.3 Our problem statement is "**To investigate women strengthening as per nations."**

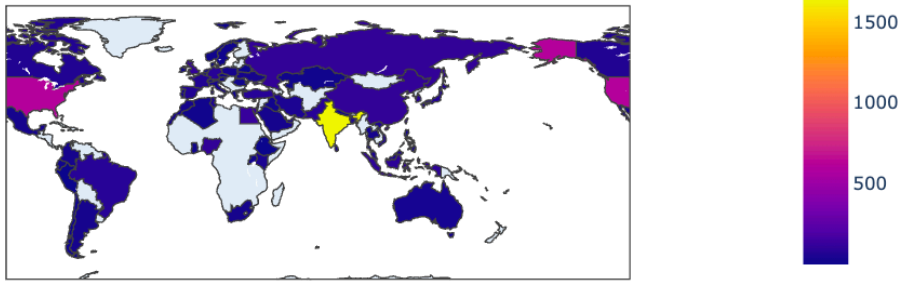



```
top_four_countries = total_woman_country.sort_values(by="Gender", ascending=False).drop("Other")[:5]
plt.figure(figsize=(12,9))
sns.barplot(x=top_four_countries.index, y=top_four_countries["Gender"])
plt.ylabel("Number Of Women Responses", fontsize=15)
plt.xlabel("Countries", fontsize=15)
plt.xticks(fontsize=15, rotation=-90)
plt.yticks(fontsize=15)
plt.show()
```



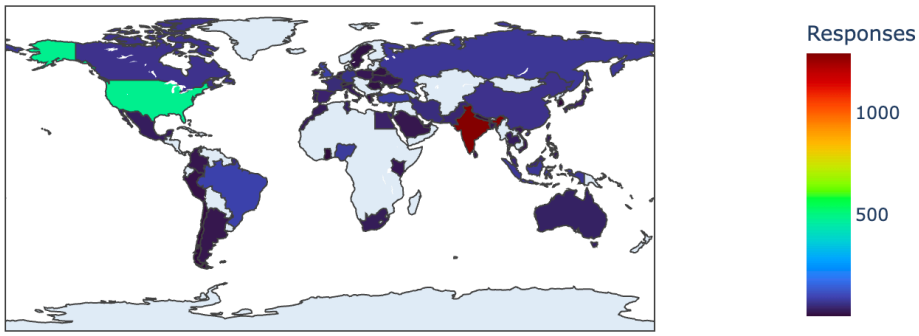
```
fig = px.choropleth(tot_woman_country_wise_df,
                    locations='Country', locationmode="country names",
                    color='Responses', color_continuous_scale=px.colors.sequential.Plasma,
                   width=800, height=400, title="Year 2021")
fig.show()
```

Year 2021




```
tot_woman_country_wise_df = pd.read_csv("r/r2020.csv")
fig = px.choropleth(tot_woman_country_wise_df,
                    locations='Country', locationmode="country names",
                    color='Responses', color_continuous_scale=px.colors.sequential.Turbo,
                   width=800, height=400, title="Year 2020")
fig.show()
```

Year 2020



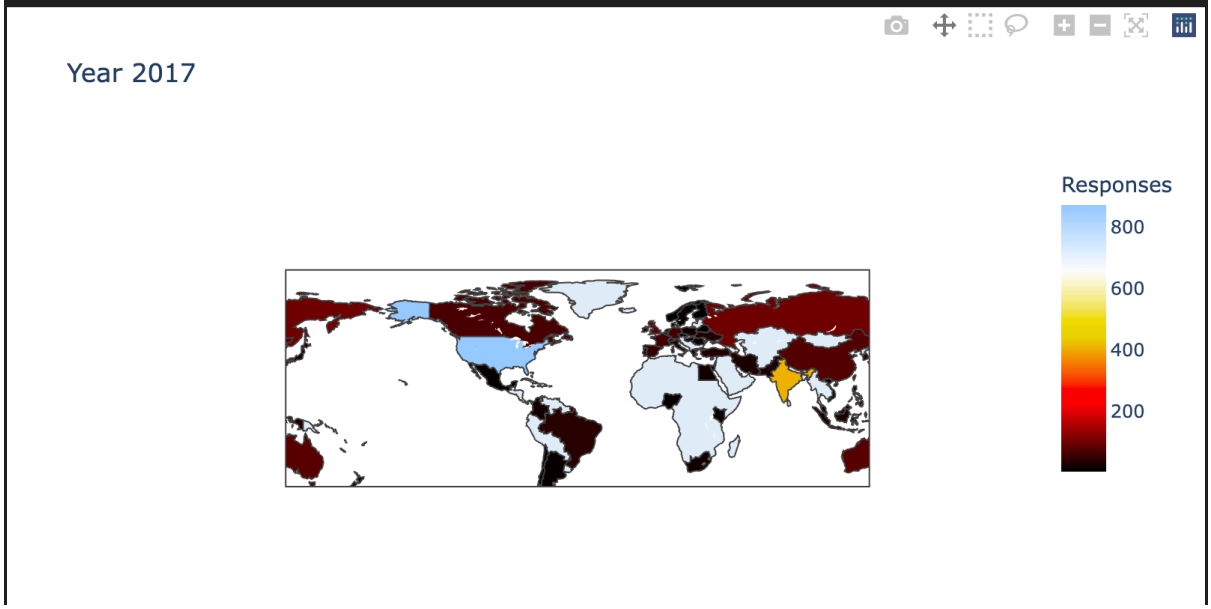
```
tot_woman_country_wise_df = pd.read_csv("r/r2019.csv")
fig = px.choropleth(tot_woman_country_wise_df,
                    locations='Country', locationmode="country names",
                    color='Responses', color_continuous_scale=px.colors.sequential.Aggrnyl,
                   width=800, height=400, title="Year 2019")
fig.show()
```

Year 2019

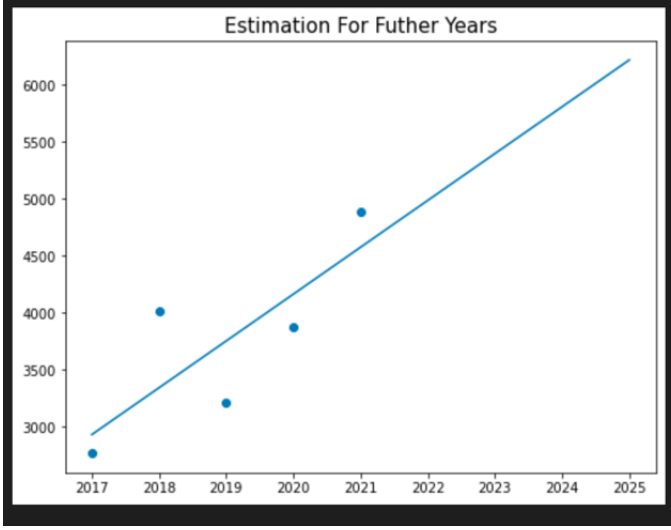

```
tot_woman_country_wise_df = pd.read_csv("r/r2018.csv")
fig = px.choropleth(tot_woman_country_wise_df,
                    locations='Country', locationmode="country names",
                    color='Responses', color_continuous_scale=px.colors.sequential.Mint,
                   width=800, height=400, title="Year 2018")
fig.show()
```

Year 2018

```
tot_woman_country_wise_df = pd.read_csv("r/r2017.csv")
fig = px.choropleth(tot_woman_country_wise_df,
                    locations='Country', locationmode="country names",
                    color='Responses', color_continuous_scale=px.colors.sequential.Blackbody,
                   width=800, height=400, title="Year 2017")
fig.show()
```




```
responses = np.array([response_2017["Responses"].sum(), response_2018["Responses"].sum(), response_2019["Responses"].sum(),
            response_2020["Responses"].sum(), response_2021["Responses"].sum()])
years = np.array([[2017], [2018], [2019], [2020], [2021]])

model = LinearRegression()
model.fit(years, responses)

esti_years = [i for i in range(2017, 2026)]
esti_respn = [int(model.predict([[year]])) for year in esti_years]

plt.figure(figsize=(8, 6))
plt.title("Estimation For Futher Years", size=15)
plt.scatter(years, responses)
plt.plot(esti_years, esti_respn)
plt.show()
```



PS 2.1 According to our Problem Statement, we are going to use **Test of Hypothesis with double proportion** and our problem statement is "**To examine the gender differences in technical employment throughout the world.** Our Statement is

"***In every field females are significantly more than men in every field.***"

Null Hypothesis à H0 : P1=P2

Alternative Hypothesis: P1≠ $P2$(Two Tailed)

$$\alpha = 0.05$$

H1=15 and H2=15

2. Test Statistics:

$$z = \frac{p_1 - p_2}{\sqrt{pq\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

Where p1; (P H1 P1 + H2 P2)/(n1+n2)

P=1-q

$|z|>z(\frac{\alpha}{2})$ (Two Tailed)

$|z|>z(0.00512)$ and $|z|>z(0.025)$

$|z|>z(\frac{\alpha}{2})=1.960$

N1=15 and N2=15

P1=0.8081 and P2=0.1918

$$z = \frac{p_1 - p_2}{\sqrt{pq\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

P=(H1 P1 + H2 P2)/(n1+n2)

$$= \frac{15 * 0.8081 + 15 * 0.1918}{30}$$

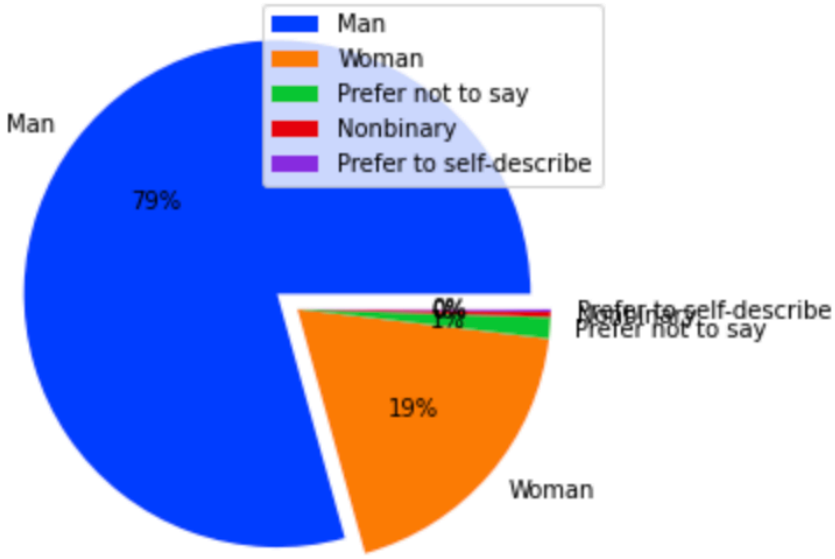
P=0.5 Q=1-Pà Q = 1-0.5à Q=0.5

$$z = \frac{0.8081 - 0.1918}{\sqrt{(0.5 * 0.5)\left(\frac{1}{15}\right) * \left(\frac{1}{15}\right)}}$$

Z=3.375

$$|z| > z\left(\frac{\alpha}{2}\right)$$

So, "***<u>In every field women are not comparatively more than men</u>***"



PS 2.2 According to our Problem Statement, we are going to use **<u>Test of Hypothesis concerning difference of means (Two Populations)</u>** and

our problem statement is "**On the basis of our data, assess and forecast the future of India's job statuses"**. Our Statement is

"***<u>In every field in Indian women are comparatively more than men"</u>***

So here are some Calculations:

$$\mu 1 = Men$$

$$\mu 2 = Women$$

Step 1: Null Hypothesis H_0: $\mu_1 - \mu_2 = d$

Alternative Hypothesis: $\mu_1 - \mu_2 > d$ (Right tailed);

$\mu_1 - \mu_2 < d$ (Left tailed);

$\mu_1 - \mu_2 =/ d$ (Two tailed)

Level of Significance:

α Sample Size: Sample 1 - n_1 and Sample 2 – n_2

¯H0=$\mu 1 - \mu 2 = 0$

Ha= $\mu 1 - \mu 2 \neq 0$

Level of Significance: $\boldsymbol{\alpha = 0.05}$

N1 = 15 and N2 = 15

N1 + N2 = 30 **(age sample $\alpha 1^2$ and $\alpha 2^2$ are not known)**

$$z = \frac{\overline{x_1} - \overline{x_2} - \delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

X1(bar)=1.4480016 and X2(bar)=0.4331449

$$\sigma 1^2 = 1.979766$$

$$\sigma 2^2 = 0.693735$$

$$\boldsymbol{z = \frac{1.4480016 - 0.4331449}{\sqrt{\left(\frac{1.979766}{15}\right) - \left(\frac{0.693735}{15}\right)}}}$$

Z=1.874

$$|z| < z\left(\frac{\alpha}{2}\right)$$

So, "***<u>In every field in Indian women are not comparatively more than men"</u>***

PS 2.3 According to our Problem Statement, we are going to use **<u>Linear Regression .</u>**and our problem statement is "**To analyse and look**

at the representative's age in the Tech Industry with male and female.”

Linear Regression is applied to predict the year in which the ratio of male and female in tech Industry is going to be approximately equal.

Sx = 891

Sy = 303

Sxy = 22257

Sx^2 = 66453

Mean(x) = 74.25

Mean(y) = 25.25

Mean(xy) = 1854.75

Mean(x^2) = 5537.75

S_{xx}:

$$S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}$$

= 66453 - (793881 / 12) = 296.25

S_{xy}:

$$S_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n}$$

= 22257-((891 * 303) / 12) = -240.57

y = bx + a

b = S_{xy}/ S_{xx} = -0.812

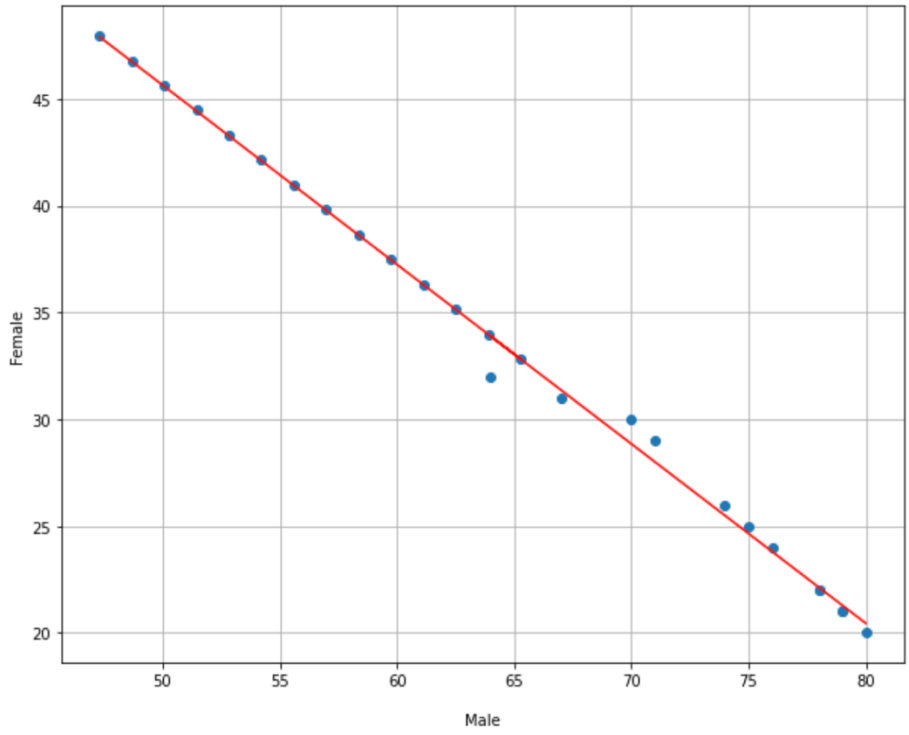
a = 85.5899

Hence, **Y = 85.5899 - 0.8127X**

Predicted Values of male and female.

2022	65	33
2023	64	34
2024	63	35
2025	61	36
2026	60	37
2027	58	39
2028	57	40
2029	56	41
2030	54	42
2031	53	43
2032	51	44
2033	50	46
2034	49	47
2035	47	47

```
In [12]: import numpy as np
         x = ps3['Male']
         y = ps3['Female']
         plt.figure(figsize=[10, 8])
         plt.plot(x, y, 'o')

         m, b = np.polyfit(x, y, 1)
         plt.xlabel('\nMale')
         plt.ylabel("Female")

         plt.plot(x, m*x + b, color='red')
         plt.grid(True)
         plt.show()
```



PS 3.1 According to our Problem Statement, we are going to use **Test of Hypothesis concerning difference of means (Two Populations)** and our problem statement is "**To concentrate on the compensations given by organizations and see the inclinations done by organizations. Our Statement is**

"The average salary for male in 15 different job category is significantly higher than female"

Null hypothesis H0

$$\mu 1 = Men$$

$$\mu 2 = Women$$

$$\mu 0: \mu 1 - \mu 2$$

Ha: $\mu 1 - \mu 2 \neq 0$

$$\alpha = 0.05$$

n1=15 and n2= 15

n1+n2=30(age sample $\sigma 1^2 \; and \; \sigma 2^2$ are not known

$$z = \frac{\overline{x_1} - \overline{x_2} - \delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Calculation x1(bar)=45481 and Calculation x2(bar)=32424

$z\left(\frac{\alpha}{2}\right) = 1.960$

$$\sigma 1^2 = 480172617$$

$$\sigma 2^2 = 280929597$$

$$\boldsymbol{z = \frac{45481 - 32424}{\sqrt{\left(\frac{480172617}{15}\right) - \left(\frac{280929517}{15}\right)}}}$$

$$\boldsymbol{z = \frac{13057}{7123}}$$

Z=1.833

$|z| < z\left(\frac{\alpha}{2}\right)$

So, ***"The average salary for male in 15 different job category is significantly higher than female."***

PS 3.2 According to our Problem Statement, we are going to use **Test of Hypothesis with double proportion** and our problem statement is and perceive how many women are doing investment into it. Our statement is

"***The average investment in tech industry, females are in 15 different job category is significantly higher than male***"

Null Hypothesis à H0 : P1=P2

Alternative Hypothesis: P1≠ $P2$(Two Tailed)

$$\alpha = 0.05$$

n1=15 and n2=15

2. Test Statistics:

$$z = \frac{\mathrm{p}_1 - p_2}{\sqrt{pq\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

Where p1; (P H1 P1 + H2 P2)/(n1+n2)

P=1-q

$|z|>z(\frac{\alpha}{2})$ (Two Tailed)

$|z|>z(0.00512)$ and $|z|>z(0.025)$

$|z|>z(\frac{\alpha}{2})=1.960$

n1=15 and n2=15

P1=0.8837 and P2=0.1163

$$z = \frac{p_1 - p_2}{\sqrt{pq\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

P=(n1 P1 + n2 P2)/(n1+n2)

$$= \frac{15 * 0.8837 + 15 * 0.1163}{30}$$

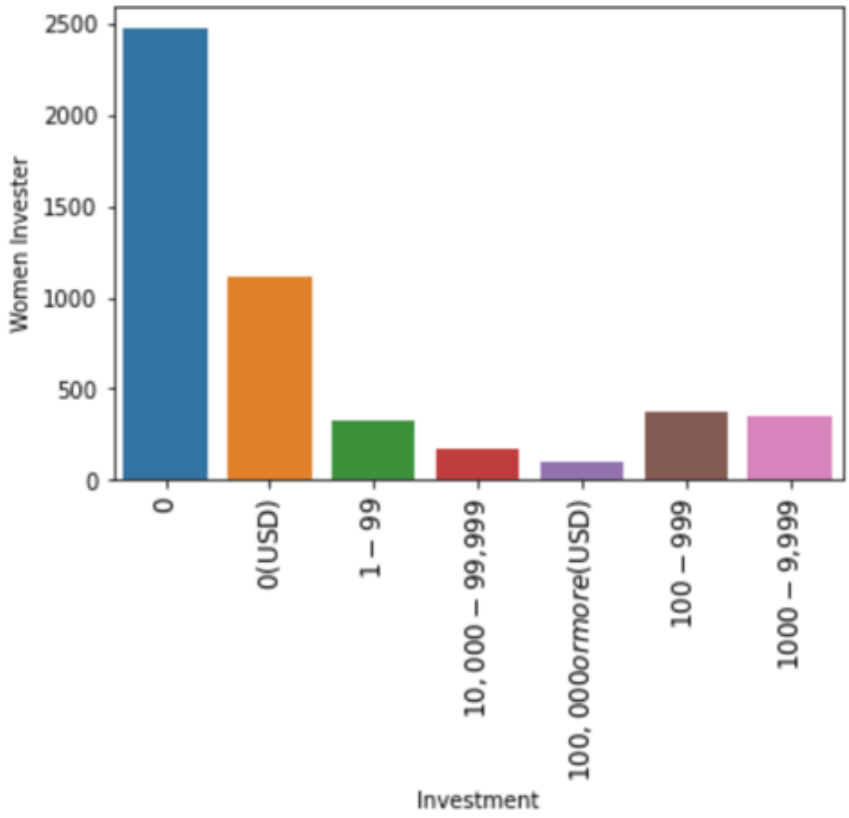
P=0.5 Q=1-Pà Q = 1-0.5à Q=0.5

$$z = \frac{0.8837 - 0.1163}{\sqrt{(0.5 * 0.5)\left(\frac{1}{15}\right) * \left(\frac{1}{15}\right)}}$$

Z=4.203

$$|z|>z(\frac{\alpha}{2})$$

So, the average investment in tech industry, females are in 15 different job category is significantly not higher than male



PS 4.1 According to our Problem Statement, we are going to use **Two variance Hypothesis** and our problem statement is "**To relate the Proportion of female and male population using the Internet around the world. And to check the claim "female users use less internet or mobile phone than boys"**

So here are some calculation:

s1=Men

s2=Women

Step 1: Null Hypothesis H_0: s1=s2

Alternative Hypothesis : s1>s2 Right tailed); : s1<s2 (Left tailed); : s1¹s2 (Two tailed) Level of Significance: α

Sample Size:

Sample 1 - n1 and Sample 2– n2

H0: s1=s2

Ha= : s1>s2

Level of Significance: $\boldsymbol{\alpha}$=$\mathbf{0.05}$

n1=18 and n2=18 (Right Tailed)

Test Statistic:

$F = \frac{S_i^2}{S_j^2}$, (where i > j)

Step 3: Criteria of Rejection

Reject the null hypothesis if F > $F_{\alpha(n_1-1)(n_2-1)}$(Right hand)

Calculation : $F = \frac{S_i^2}{S_j^2} = \frac{(174.247)^2}{(147.975)^2}$ = 1.386

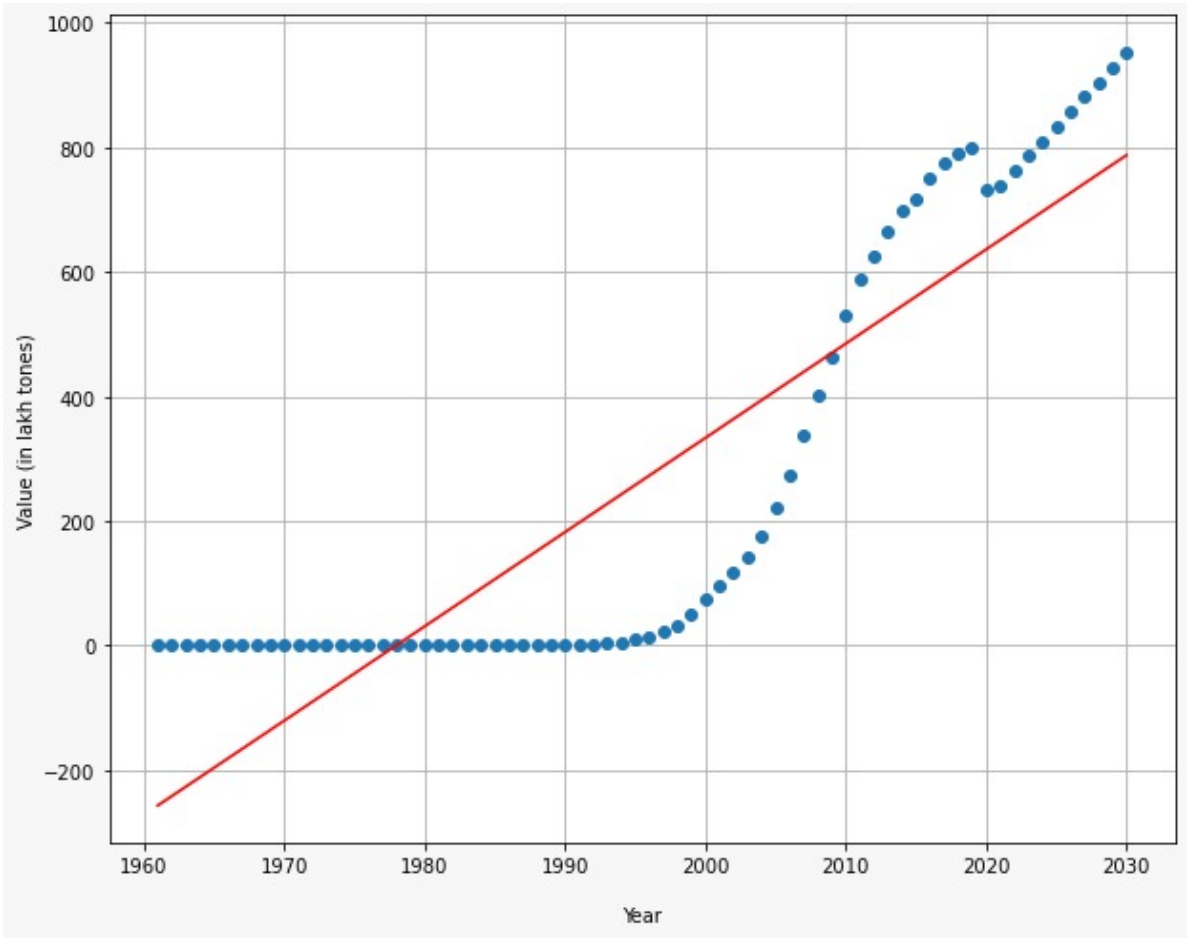
$F_{\alpha(n_1-1)(n_2-1)}$=$F_{(0.5*17*17)}$ *= 2.31(according to table)*

F > $F_{\alpha(n_1-1)(n_2-1)}$: 1.386<2.31(Right Hand)

So, "Female users use less internet or mobile phone than boys."

PS 4.2 According to our Problem Statement, we are going to use **Linear Regression** and our problem statement is "**To predict total number cellular data user for next decade based on current data.**"

```
from sklearn import linear_model
reg = linear_model.LinearRegression()
reg.fit(nndf[['Year']],nndf['Mobile Access'])

num1 = np.array([i for i in range(2021, 2031)])
year = pd.DataFrame(num1)
predicted_vals = regC.predict(year)
predicted_vals

x = list(ndf['Year']) + list(num1)
y = list(ndf['Mobile Access']) + list(pr)

final_df = pd.DataFrame()
final_df['Year'] = x
final_df['Mobile Access'] = y
final_df

x = final_df['Year']
y = final_df['Mobile Access']
plt.figure(figsize=[10, 8])
plt.plot(x, y, 'o')

m, b = np.polyfit(x, y, 1)
plt.xlabel('\nYear')
plt.ylabel("Value (in lakh tones)")

plt.plot(x, m*x + b, color='red')
plt.grid(True)
plt.show()
```

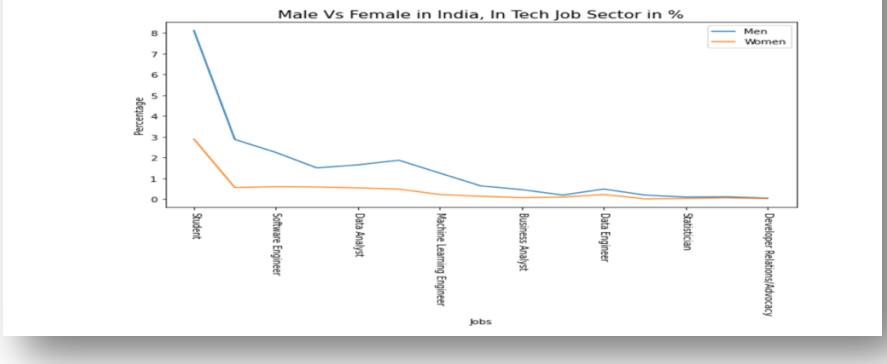


Result :

As this graph shows in 2019 due to corona virus there is a significant difference and we have predicted cellular data user for a decade.

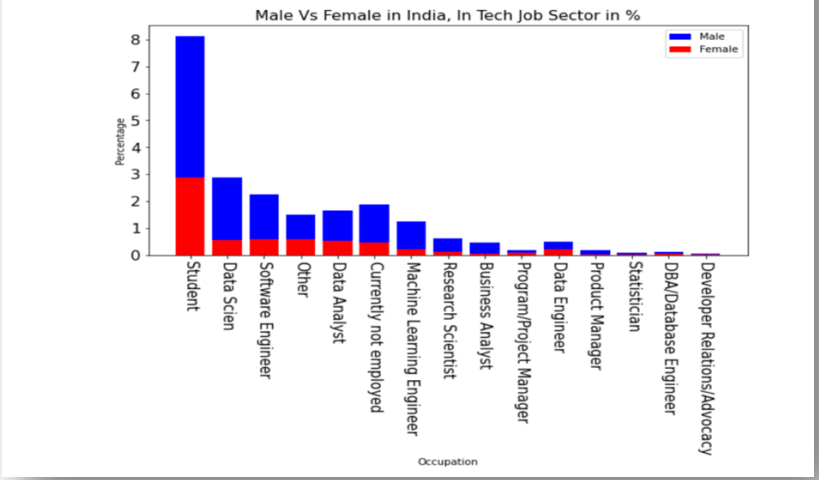
Graph

```
In [82]: import matplotlib.pyplot as plt
         df=pd.read_csv("ps22.csv")
         df=df.set_index(df['Age '])
         df.plot(figsize=(10,5))
         plt.xlabel("Jobs")
         plt.ylabel("Percentage")
         plt.title("Male Vs Female in India, In Tech Job Sector in %", size=15)
         plt.xticks(rotation=-90, size=10)

Out[82]: (array([-2.,  0.,  2.,  4.,  6.,  8., 10., 12., 14., 16.]),
          [Text(-2.0, 0, 'DBA/Database Engineer'),
           Text(0.0, 0, 'Student'),
           Text(2.0, 0, 'Software Engineer'),
           Text(4.0, 0, 'Data Analyst'),
           Text(6.0, 0, 'Machine Learning Engineer'),
           Text(8.0, 0, 'Business Analyst'),
           Text(10.0, 0, 'Data Engineer'),
           Text(12.0, 0, 'Statistician'),
           Text(14.0, 0, 'Developer Relations/Advocacy'),
           Text(16.0, 0, '')])
```

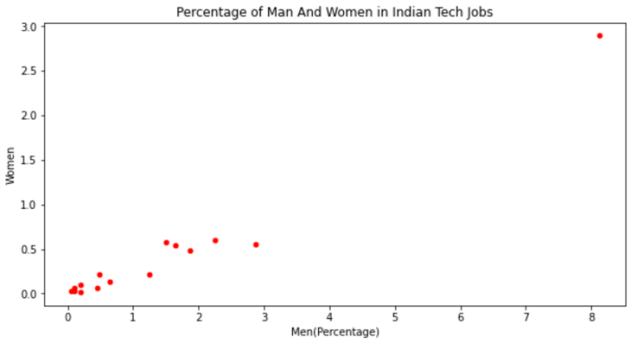


Male and Female in Indian Tech Jobs

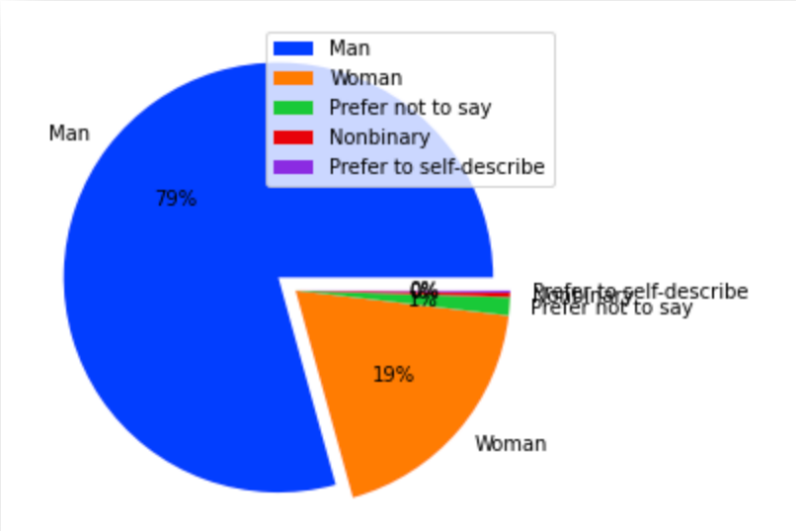
```
In [30]: df=df.set_index(df['Age '])
         plt.figure (figsize= (10, 5))
         plt.bar(df.index, df["Men "], label= 'Male',color="blue")
         plt.bar(df.index, df["Women"], label = 'Female',color="red")
         plt.title("Male Vs Female in India, In Tech Job Sector in %", size=15)
         plt.xlabel("Occupation")
         plt.ylabel("Percentage")
         plt.xticks (rotation=-90, size=15)
         plt.yticks(size=15)
         plt.legend()
         plt.show()
```



```
In [89]: ])
='Men ',y='Women',figsize=(10,5),color="red",title="Percentage of Man And Women in Indian Tech Jobs",xlabel="Men(Percen

Out[89]: <AxesSubplot:title={'center':'Percentage of Man And Women in Indian Tech Jobs'}, xlabel='Men(Percentage)', ylabel='Wo
men'>
```

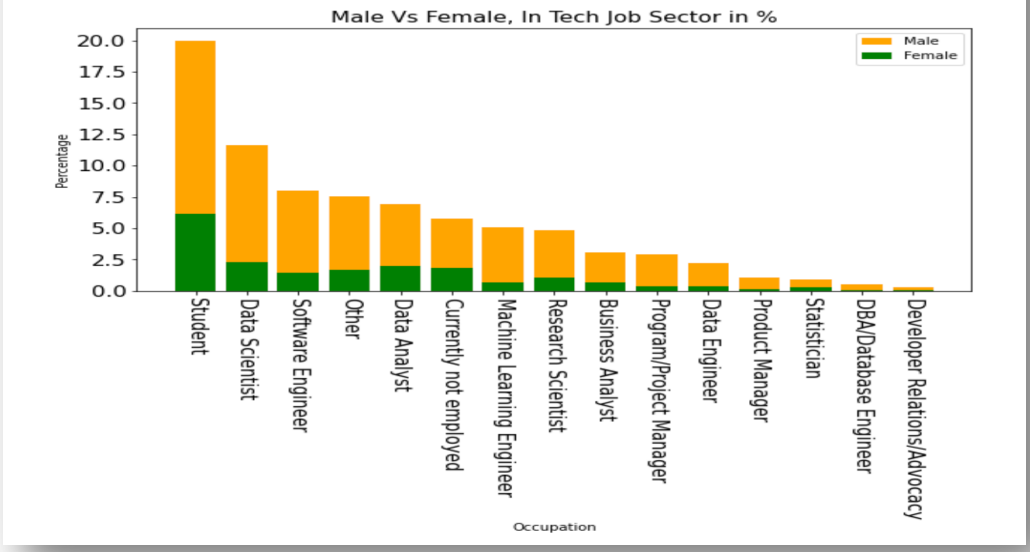


Scatter graph of male and female in Indian tech sector.



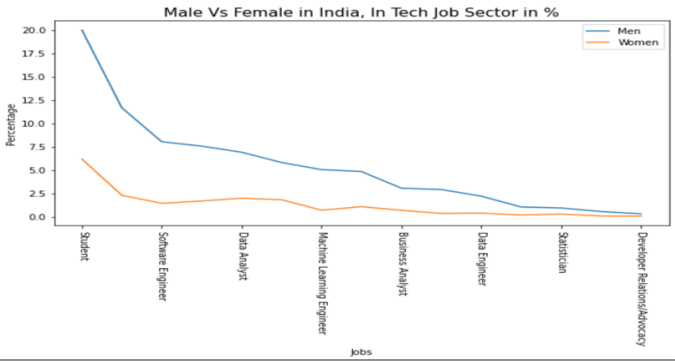
Pie Chart male and female jobs in World

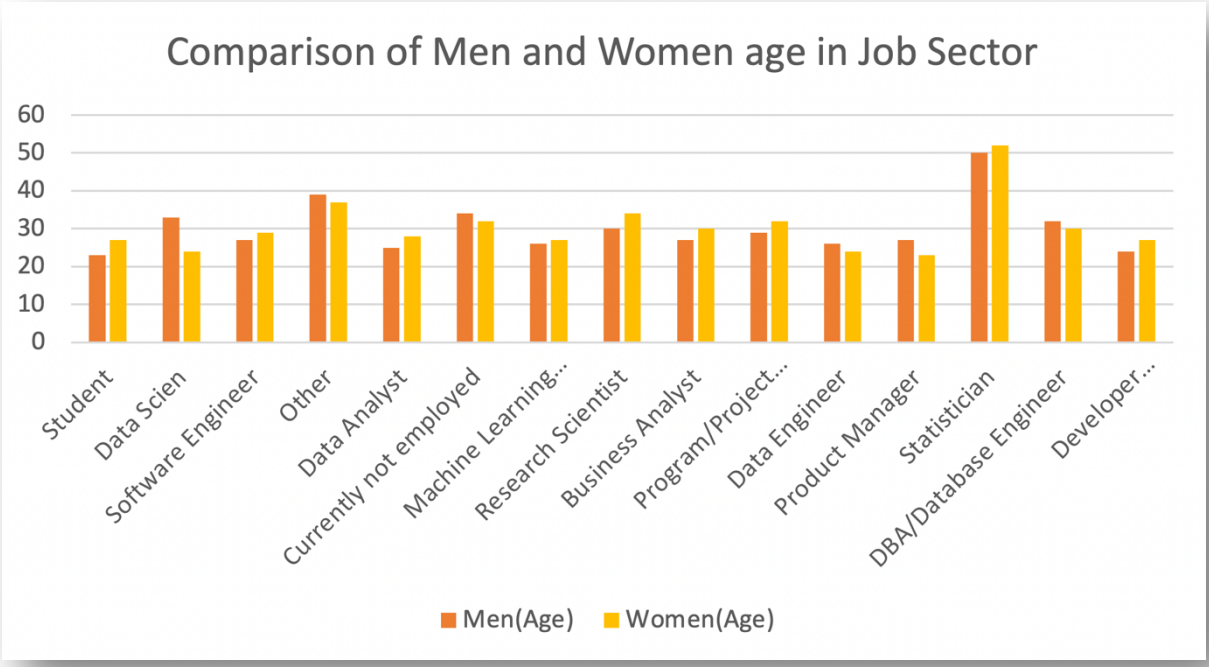
```
]: df2=df2.set_index(df2['Occupation'])
   plt.figure(figsize= (10, 5))
   plt.bar(df2.index, df2["Men "], label= 'Male',color="orange")
   plt.bar(df2.index, df2["Women"], label = 'Female',color="green")
   plt.title("Male Vs Female, In Tech Job Sector in %", size=15)
   plt.xlabel("Occupation")
   plt.ylabel("Percentage")
   plt.xticks (rotation=-90, size=15)
   plt.yticks(size=15)
   plt.legend()
   plt.show()
```



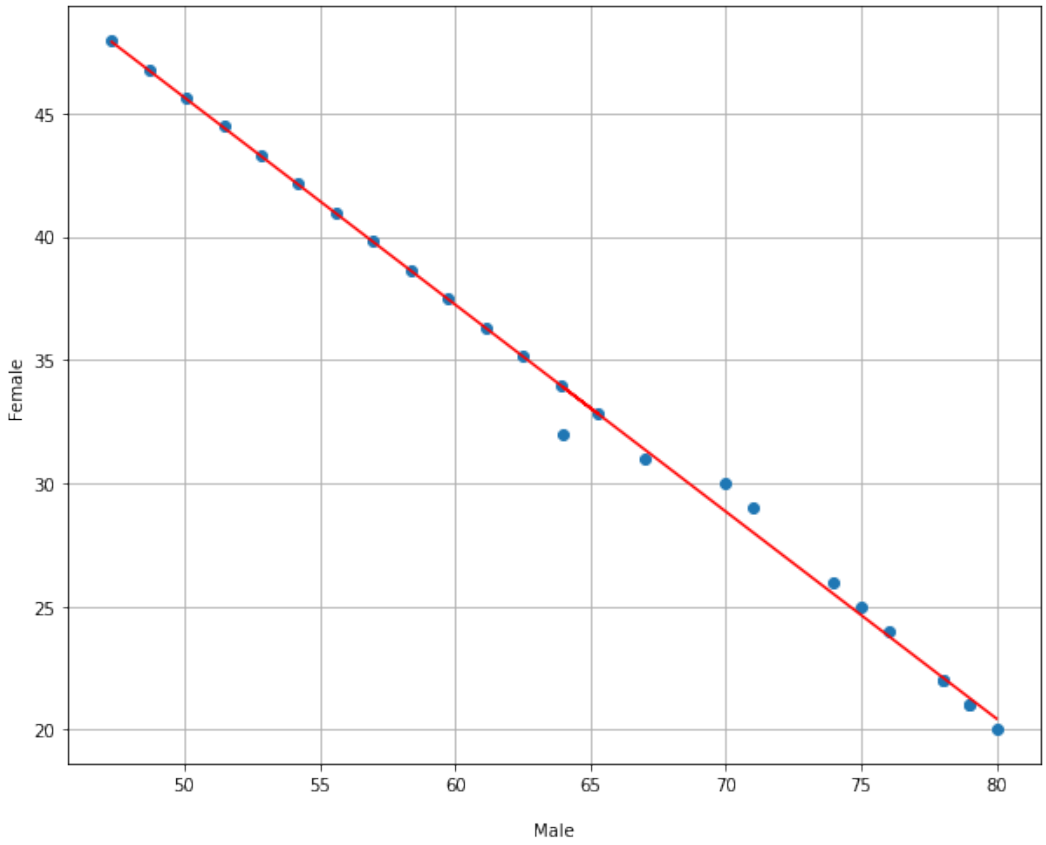
Male and Female in Job Sector

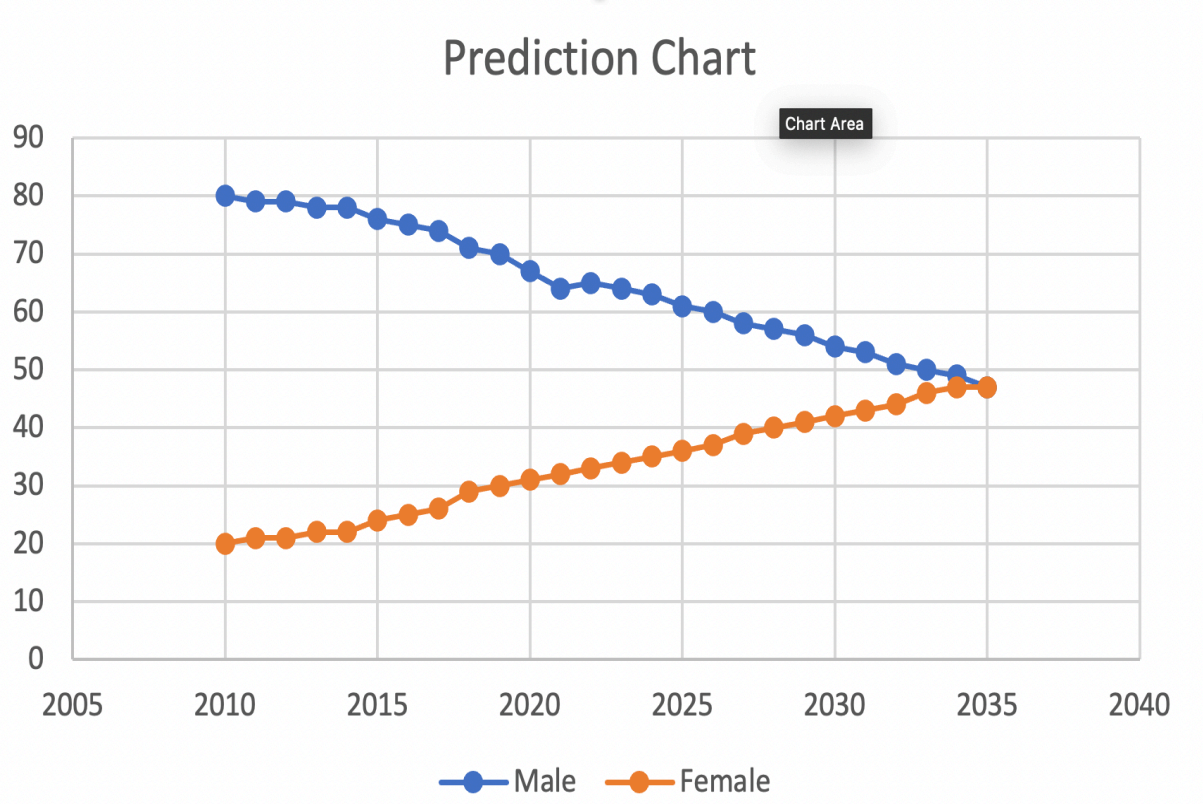
```
In [103]: import matplotlib.pyplot as plt
          d2f=pd.read_csv("ps21.csv")
          df2=df2.set_index(df2['Occupation'])
          df2.plot(figsize=(10,5))
          plt.xlabel("Jobs")
          plt.ylabel("Percentage")
          plt.title("Male Vs Female, In Tech Job Sector in %", size=15)
          plt.xticks(rotation=-90, size=10)

Out[103]: (array([-2.,  0.,  2.,  4.,  6.,  8., 10., 12., 14., 16.]),
           [Text(-2.0, 0, 'DBA/Database Engineer'),
            Text(0.0, 0, 'Student'),
            Text(2.0, 0, 'Software Engineer'),
            Text(4.0, 0, 'Data Analyst'),
            Text(6.0, 0, 'Machine Learning Engineer'),
            Text(8.0, 0, 'Business Analyst'),
            Text(10.0, 0, 'Data Engineer'),
            Text(12.0, 0, 'Statistician'),
            Text(14.0, 0, 'Developer Relations/Advocacy'),
            Text(16.0, 0, '')])
```





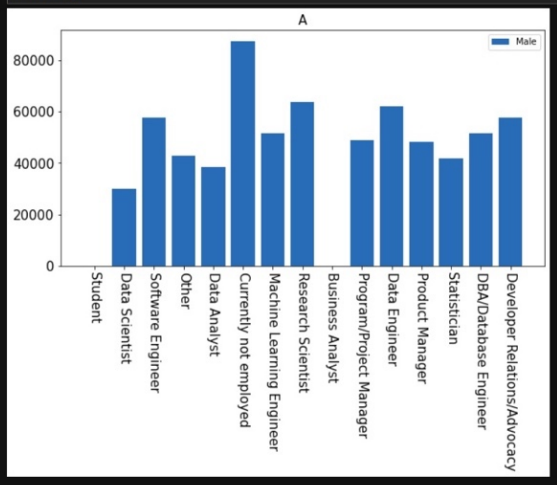
Linear Regression graph



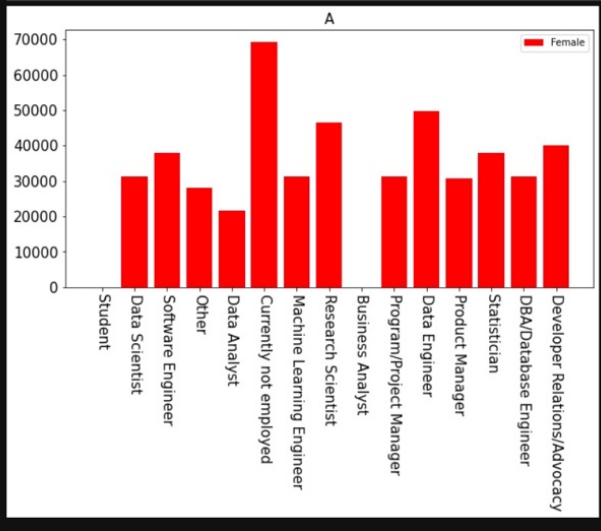


This Scatter Graph shows the prediction values of jobs, which is predicted that female are going to cross male in 2035.

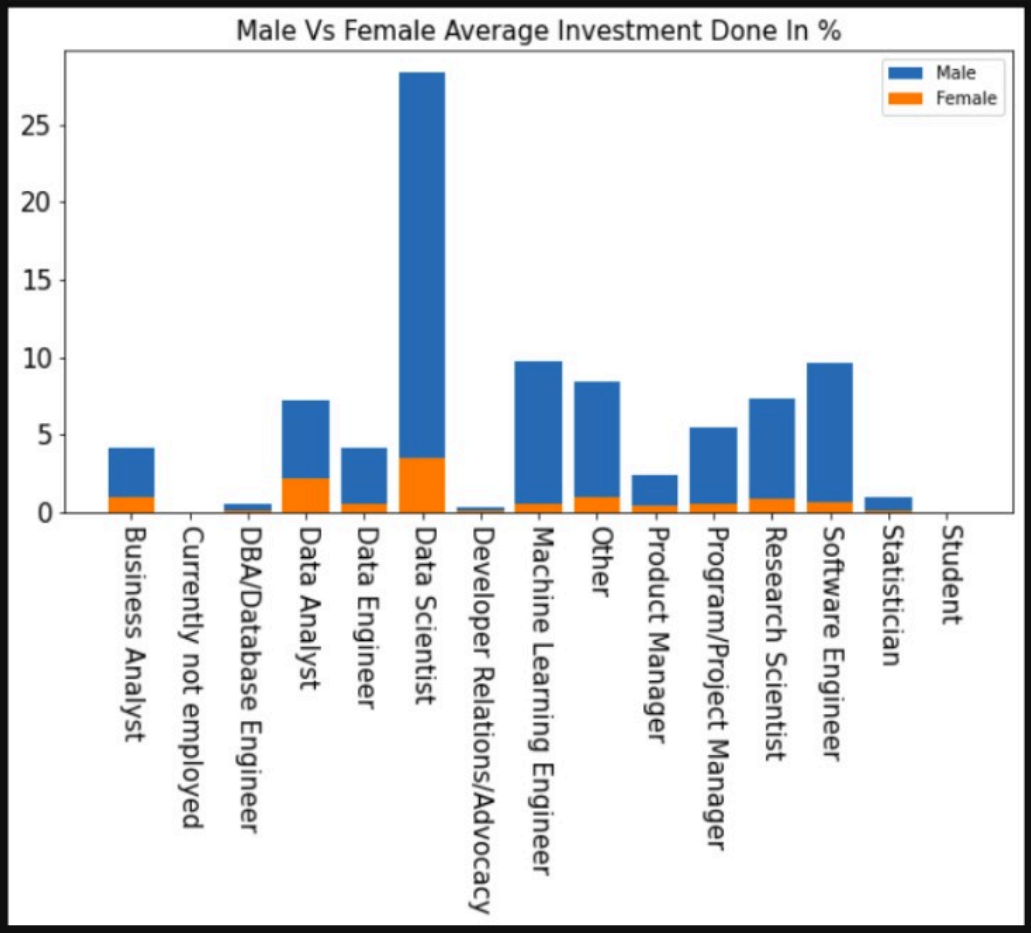
```
[8]: plt.figure(figsize=(10, 5))
     plt.bar(kdf.index, kdf["Men "], label = 'Male')
     # plt.bar(kdf.index, kdf["Women"], label = 'Female')
     plt.title("A", size=15)
     plt.xticks(rotation=-90, size=15)
     plt.yticks(size=15)
     plt.legend()
     plt.show()
```

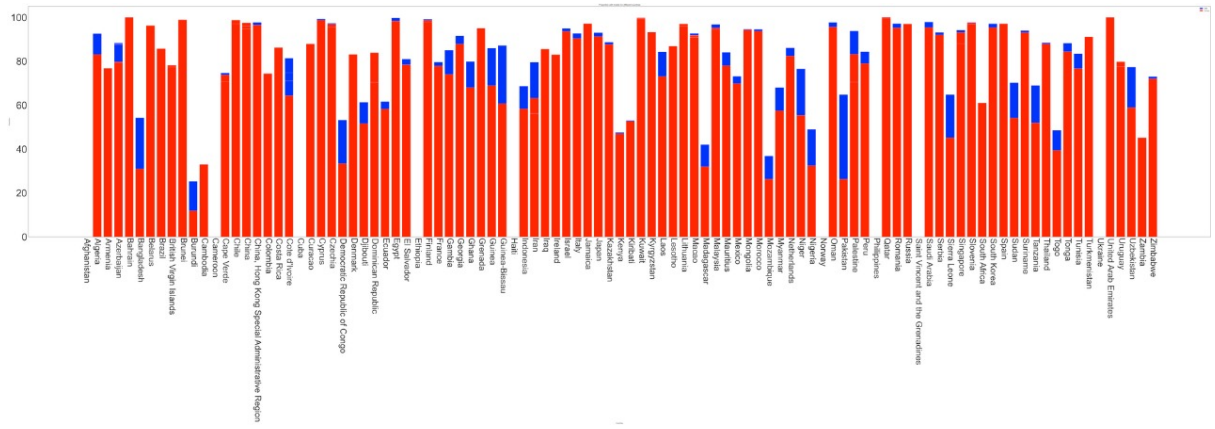



```
1]: plt.figure(figsize=(10, 5))
    # plt.bar(kdf.index, kdf["Men "], label = 'Male')
    plt.bar(kdf.index, kdf["Women"], label = 'Female', color="red")
    plt.title("A", size=15)
    plt.xticks(rotation=-90, size=15)
    plt.yticks(size=15)
    plt.legend()
    plt.show()
```

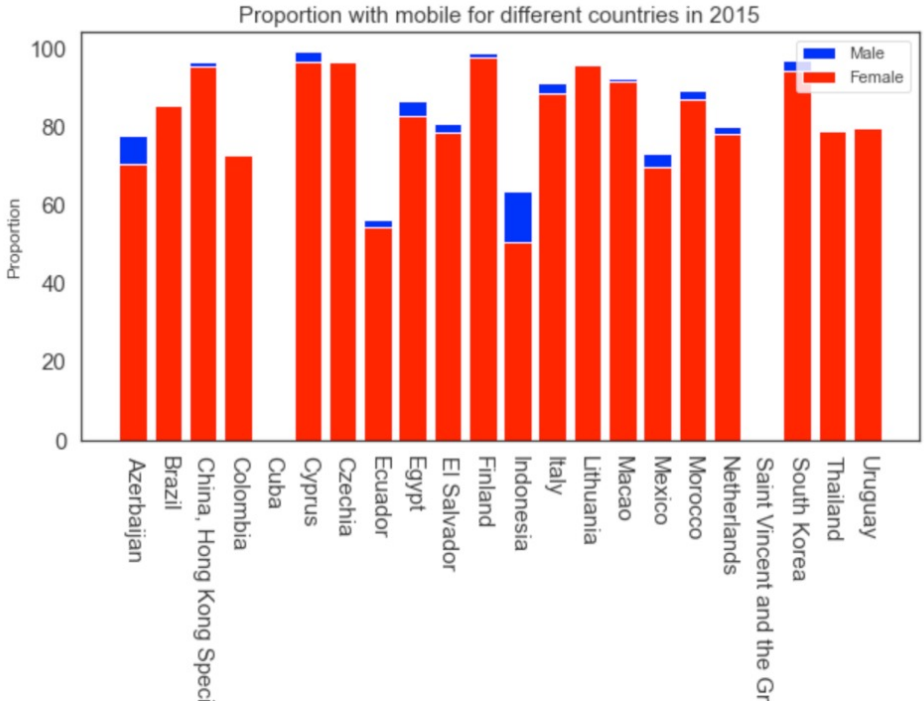


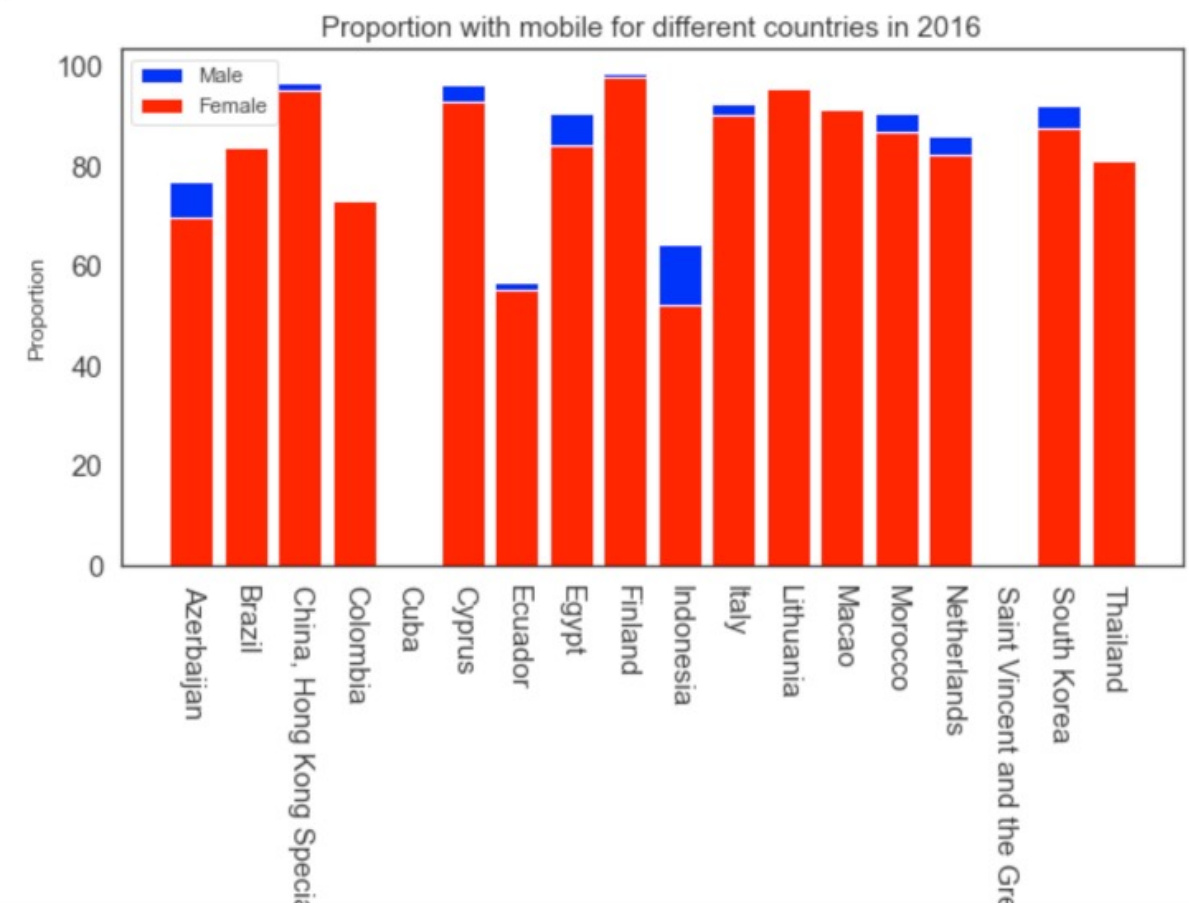
```
27]: plt.figure(figsize=(10, 5))
     plt.bar(per_df.index, per_df["Male"], label = 'Male')
     plt.bar(per_df.index, per_df["Female"], label = 'Female')
     plt.title("Male Vs Female Average Investment Done In %", size=15)
     plt.xticks(rotation=-90, size=15)
     plt.yticks(size=15)
     plt.legend()
     plt.show()
```

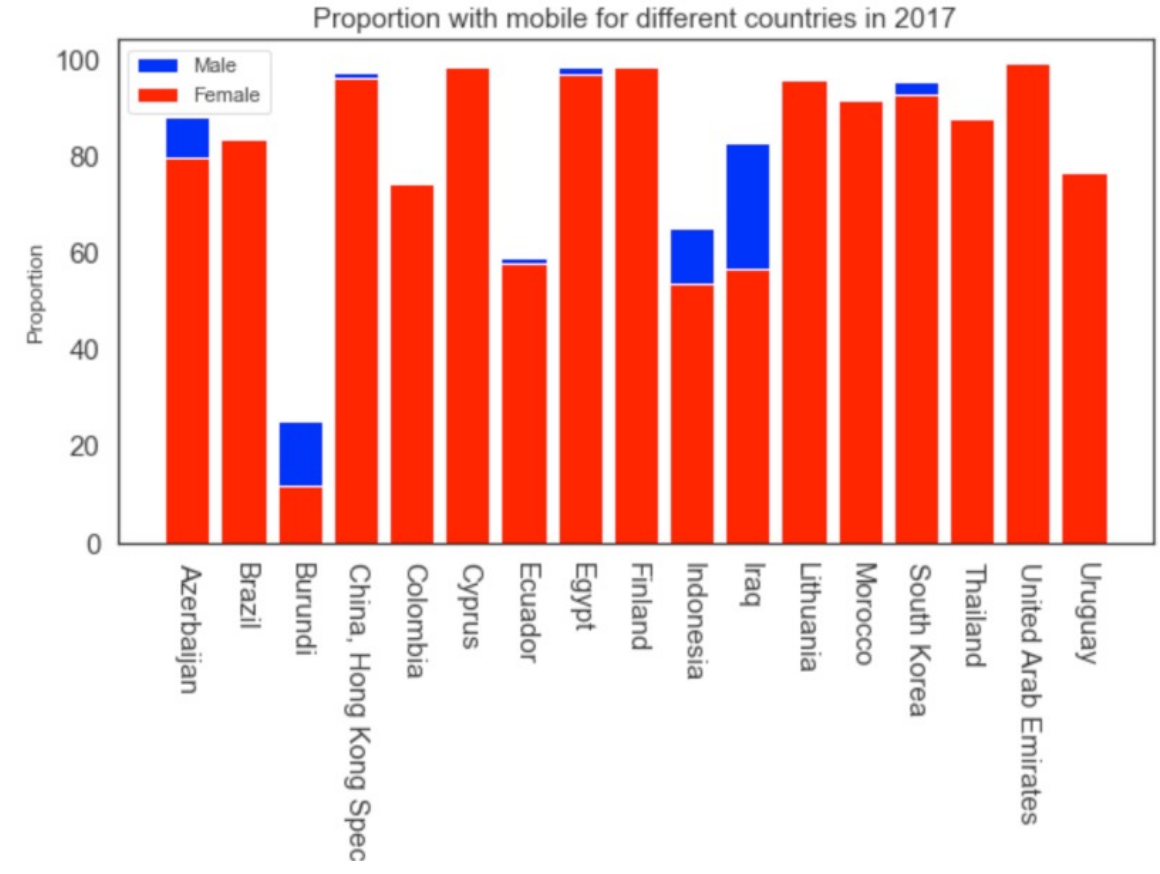


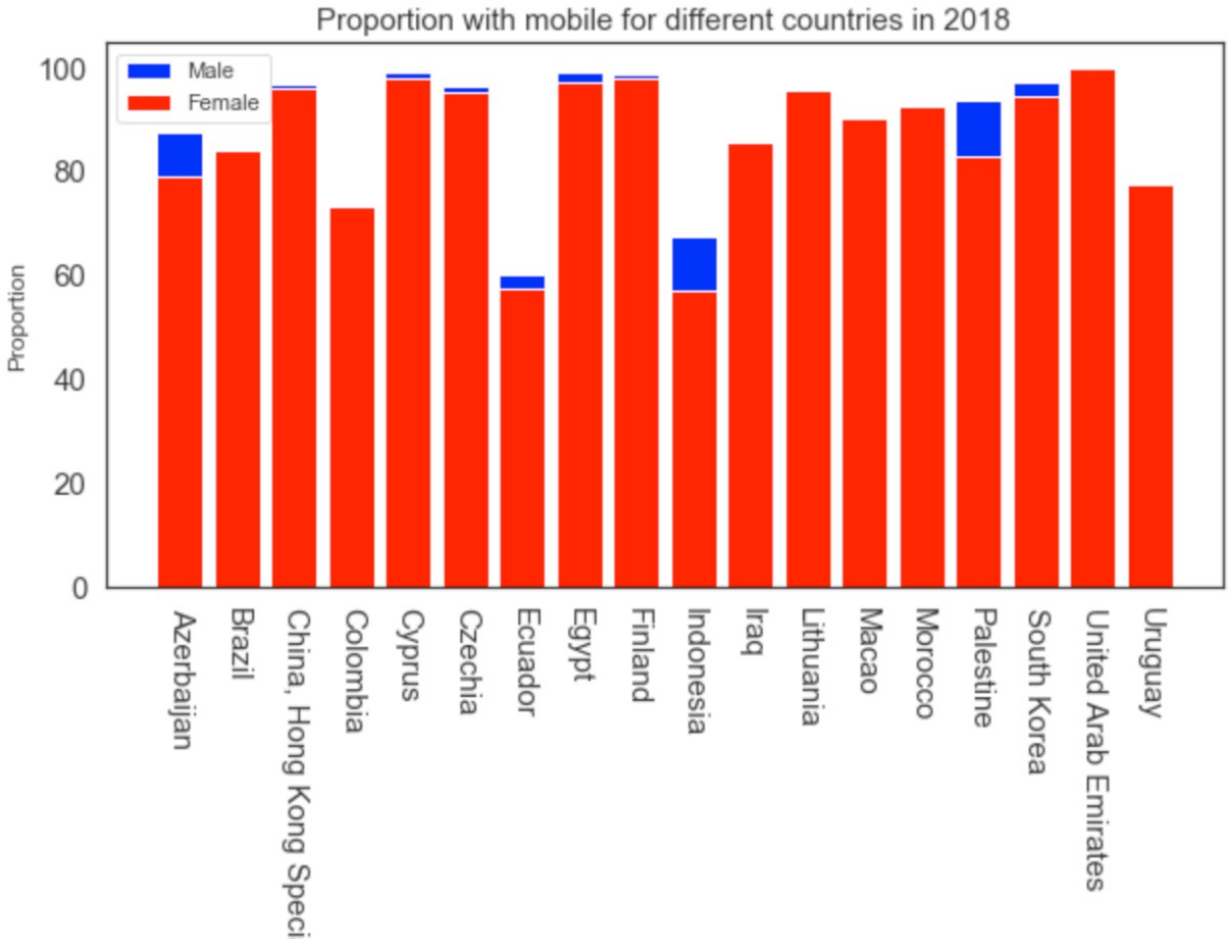


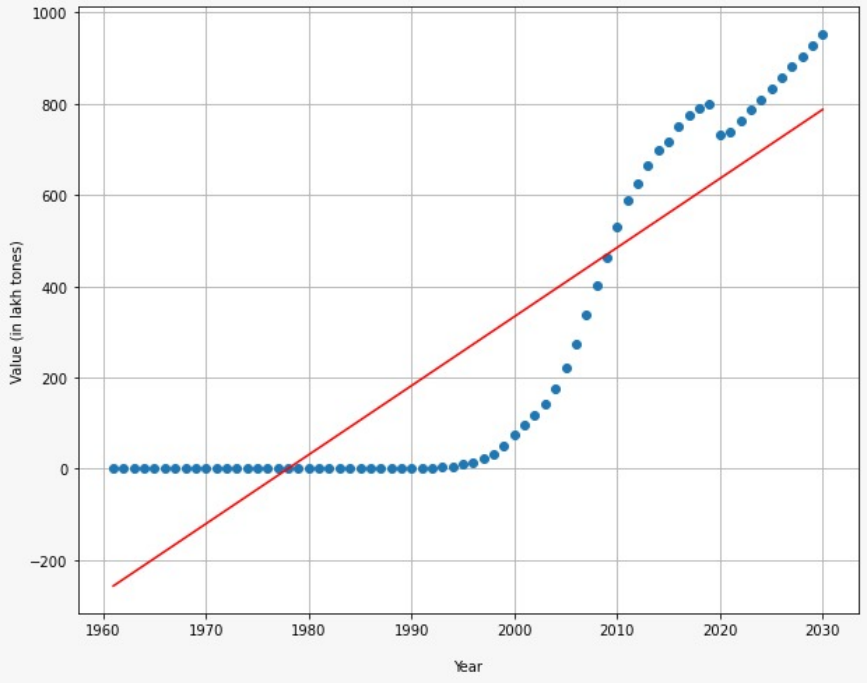
```
plt.figure(figsize= (10, 5))
plt.bar(df15['Country'], df15["Proportion with mobile(%) - male"], label= 'Male',color="blue")
plt.bar(df15['Country'], df15["Proportion with mobile(%) - female"], label = 'Female',color="red")
plt.title("Proportion with mobile for different countries in 2015", size=15)
plt.xlabel("Countries")
plt.ylabel("Proportion")
plt.xticks (rotation=-90, size=15)
plt.yticks(size=15)
plt.legend()
plt.show()
```




Proportion with mobile for different countries in 2016
Male
Female
Proportion
100
80
60
40
20
0
Azerbaijan
Brazil
China, Hong Kong Speci
Colombia
Cuba
Cyprus
Ecuador
Egypt
Finland
Indonesia
Italy
Lithuania
Macao
Morocco
Netherlands
Saint Vincent and the Gr
South Korea
Thailand


Proportion with mobile for different countries in 2017
Male
Female
Proportion
100
80
60
40
20
0
Azerbaijan
Brazil
Burundi
China, Hong Kong Spec
Colombia
Cyprus
Ecuador
Egypt
Finland
Indonesia
Iraq
Lithuania
Morocco
South Korea
Thailand
United Arab Emirates
Uruguay


Proportion with mobile for different countries in 2018
Male
Female
Proportion
100
80
60
40
20
0
Azerbaijan
Brazil
China, Hong Kong Speci
Colombia
Cyprus
Czechia
Ecuador
Egypt
Finland
Indonesia
Iraq
Lithuania
Macao
Morocco
Palestine
South Korea
United Arab Emirates
Uruguay


1000
800
600
400
200
0
−200
Value (in lakh tones)
1960
1970
1980
1990
2000
2010
2020
2030
Year

Result and Conclusion

After seeing the result of problem statement's, where we use the hypothesis and some other mathematical tools like regression correlation. We see that our all hypothesis statements were rejected and there is no correlation between men and women in tech jobs. So the final result that we get is We need Women in our tech industry to compete with men and need to grow.

We see that, Males are more than females in every field of tech. Rather than tech, in different sectors females are lesser and males are more in different fields

Gender equity stays a significant issue in the corporate world, and female remain fundamentally underrepresented in the corporate pipeline. Despite a wealth of examination affirming that organizations are more productive when they have more female in the C-suite, we have an orientation hole in many organizations. Variety and consideration can't be essential for a one-time crusade; rather, they are causes that require constant work that should be created, kept up with and developed.

We see that Males are more than females in every field of tech. Rather than tech, in different sectors, females are lesser and males are more in different fields. So we saw that the result of one of the problems is - there is no correlation b/w male and female in skills set. We saw some graphs and some maps also we saw that the growth in skill set in previous years females are doing great.

It is the same in the Jobs sector also you see that in objective 2 results. Males are significantly high in every field of job if we talk about India the result is the same. we cannot say these females are not growing, in previous years the growth rate of females in jobs are increasing and we can say that by the result of problem statement 3 of objective 2: if the growth rate is the same for females in the tech industry, Females may beat Males before 2035 and in 2035 we may saw and say females are more than males in Tech jobs.

So our Conclusion is females are growing at some good rates and they have to grow more to beat males in the tech industry before 2035.

In our Objective 4 we saw that there is a huge difference due to Covid-19 which affects our analyses. In this Report we have predicted cellular data user of a decade which shows there is an exponential increase in use of mobile phones and devices.

Reference

- https://iccwbo.org/media-wall/news-speeches/3-reasons-ict-matters-gender-equality/ (12-01-2022)
- https://stats.unctad.org/Dgff2016/people/goal5/target_5_b.html (13-01-2022)
- https://sdgs.un.org/goals (15-01-2022)
- https://www.kaggle.com/kaggle -survey-2021 (11-02-2022)
- https://www.geeksforgeeks.org/python-programming-language/?ref=shm (1-01-2022)
- https://simplypsychology.org/what-is-a-hypotheses.html (12-01-2022)
- https://www.unwomen.org/en/news/stories/2021/10/feature-what-does-gender-equality-look-like-today (21-01-2022)
- https://www.un.org/en/global-issues/gender-equality (17-01-2022)
- https://hbr.org/2019/10/why-techs-approach-to-fixing-its-gender-inequality-isnt-working (12-01-2022)
- https://unstats.un.org/sdgs/unsdg (26-01-2022)
- https://www.javatpoint.com/python-tutorial (28-01-2022)
- https://sdg-tracker.org/gender-equality (30-01-2022)
- https://data.worldbank.org/indicator/IT.CEL.SETS?end=2020&start=1960 (31-01-2022)

S.NO	Name of Students with Roll no.	Contribution	Signature
1.	**Abhishek Saini (2021btech003)**	Data Collection ,Data Understanding, Data Cleaning and Objective 1	
2.	**Abhishek Swami (2021btech004)**	Data Collection ,Graph, Data Cleaning and Objective 4	
3.	**Anurag Sharma (2021btech021)**	Graph ,Report Writing , Data Cleaning and Objective 2	
4.	**Kapil Saini (2021btech056)**	Graph, Data Manipulation, Methodology and Objective 3	

Appendix

```
import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

%matplotlib inline

RAW_DATASET_LOC = "Obj2_Dataset/kaggle_survey_2021_responses.csv"

raw_dataset_df = pd.read_csv(RAW_DATASET_LOC, low_memory=False, skiprows=[0])

raw_dataset_df = raw_dataset_df.iloc[:, :232]

raw_dataset_df.head()

raw_dataset_df.rename(columns={'Duration (in seconds)': 'Response Time',

'What is your age (# years)?': 'Age',

'What is your gender? - Selected Choice': 'Gender',

'In which country do you currently reside?': 'Country',

'What is the highest level of formal education that you have attained or plan to
attain within the next 2 years?': 'Education',

'Select the title most similar to your current role (or most recent title if retired):
- Selected Choice': 'Job Title',

'For how many years have you been writing code and/or programming?': 'Experience'},
inplace=True, errors='raise')

raw_dataset_df

simplified_columns = {i:i.split("-")[-1].strip() for i in raw_dataset_df.iloc[:,
7:].columns}

raw_dataset_df.rename(columns=simplified_columns, inplace=True)

raw_dataset_df

columns_to_drop = [

'None',

'MATLAB',

'Alteryx',

'Other',

'3, etc)',
```

```
'No / None',
'Approximately how many times have you used a TPU (tensor processing unit)?',
'For how many years have you used machine learning methods?',
'CNN, etc)',
'3, BERT, XLnet, etc)',
'Selected Choice',
'What is the size of the company where you are employed?',
'Approximately how many individuals are responsible for data science workloads at your place of business?',
'Does your current employer incorporate machine learning methods into their business?',
'Analyze and understand data to influence product or business decisions',
'Build and/or run the data infrastructure that my business uses for storing, analyzing, and operationalizing data',
'Build prototypes to explore applying machine learning to new areas',
'Build and/or run a machine learning service that operationally improves my product or workflows',
'Experimentation and iteration to improve existing ML models',
'Do research that advances the state of the art of machine learning',
'None of these activities are an important part of my role at work',
'What is your current yearly compensation (approximate $USD)?',
'Approximately how much money have you (or your team) spent on machine learning and/or cloud computing services at home (or at work) in the past 5 years (approximate $USD)?',
]
raw_dataset_df.drop(columns_to_drop, axis=1, inplace=True)
raw_dataset_df
raw_dataset_df_1 = raw_dataset_df.iloc[:, :7]
raw_dataset_df_2 = raw_dataset_df.iloc[:, 7:]
raw_dataset_df_2_cols = raw_dataset_df_2.columns
raw_dataset_df_2.fillna(0, inplace=True)
raw_dataset_df_2 = (raw_dataset_df_2[raw_dataset_df_2_cols] != 0).astype(int)
dataset = pd.concat([raw_dataset_df_1, raw_dataset_df_2], axis=1)
dataset.to_csv("Obj2_Dataset/New_Ra-Do-Woo's Dataset.csv")
gender = dataset['Gender']
```

```
experience = dataset['Experience']
ps_one_df = pd.concat([gender, experience], axis=1)
ps_one_df.groupby("Experience").sum()
skill_set = dataset.iloc[:, 7:].sum(axis=1)
gender = dataset['Gender']
ps_one_df = pd.concat([gender, skill_set], axis=1)
ps_one_df.columns = ["Gender", "Skill Set"]
ps_one_df = ps_one_df.groupby("Gender").sum()
ps_one_df.sort_values("Skill Set", ascending=False, inplace=True)
ps_one_df
plt.figure(figsize=(10,5))
sns.barplot(x=ps_one_df.index, y=ps_one_df["Skill Set"])
plt.xticks(fontsize=12)
plt.yticks(fontsize=12)
plt.show()
women_dataset = dataset.loc[dataset['Gender'] == "Woman"]
skill_set = women_dataset.iloc[:, 7:].sum(axis=1)
country = dataset['Country']
ps_three_df = pd.concat([country, skill_set], axis=1, ignore_index=True)
ps_three_df.columns = ["Country", "Skill Set"]
ps_three_df = ps_three_df.groupby("Country").sum()
ps_three_df.sort_values("Skill Set", ascending=False, inplace=True)
# ps_three_df
plt.figure(figsize=(10,5))
sns.barplot(x=ps_one_df.index, y=ps_one_df["Skill Set"])
plt.xticks(fontsize=12)
plt.yticks(fontsize=12)
plt.show()
plt.figure(figsize=(15,10))
sns.barplot(x=ps_three_df.index, y=ps_three_df["Skill Set"])
plt.xticks(fontsize=12, rotation=90)
plt.yticks(fontsize=12)
```

```
plt.show()


women_dataset = dataset.loc[dataset['Gender'] == "Man"]
skill_set = women_dataset.iloc[:, 7:].sum(axis=1)
country = dataset['Country']
ps_three_df = pd.concat([country, skill_set], axis=1, ignore_index=True)
ps_three_df.columns = ["Country", "Skill Set"]
ps_three_df = ps_three_df.groupby("Country").sum()
ps_three_df.sort_values("Skill Set", ascending=False, inplace=True)
plt.figure(figsize=(15,10))
sns.barplot(x=ps_three_df.index, y=ps_three_df["Skill Set"])
plt.xticks(fontsize=12, rotation=90)
plt.yticks(fontsize=12)
plt.show()


dataset = pd.read_csv("Obj2_Dataset/New_Ra-Do-Woo's Dataset.csv")
dataset.drop("Unnamed: 0", inplace=True, axis=1)
dataset.head()


adf = (dataset.loc[dataset["Gender"] == "Woman"]).replace("Woman",
1).groupby("Country").sum()
plt.figure(figsize=(15,10))
sns.barplot(x=adf.index, y=adf["Gender"])
plt.xticks(fontsize=12, rotation=90)
plt.yticks(fontsize=12)
plt.show()


dataset = pd.read_csv("Obj2_Dataset/New_Ra-Do-Woo's Dataset.csv")
dataset.drop("Unnamed: 0", inplace=True, axis=1)
dataset.head()


response_count = dataset["Gender"].value_counts()
```

```
plt.figure(figsize=(10,5))
colors = sns.color_palette('bright')
plt.pie(response_count, labels = response_count.index, colors = colors, autopct='%.0f%%', explode=[0.1, 0, 0, 0, 0])
plt.legend()
plt.show()

response_count = dataset["Age"].value_counts()
plt.figure(figsize=(10,5))
explode = [0.19, 0.05, 0.05, 0, 0, 0, 0, 0, 0, 0, 0]
plt.pie(response_count, labels = response_count.index, colors = colors, autopct='%.0f%%', explode=explode)
plt.show()
raw_dataset_df = pd.read_csv(RAW_DATASET_LOC, low_memory=False, skiprows=[0])
raw_dataset_df.head()

investment = raw_dataset_df.loc[dataset['Gender'] == "Woman"]["Investment"]
investment.fillna(0, inplace=True)
gender = (raw_dataset_df.loc[investment.index]["Gender"] == "Woman").astype(int)

temp_df = pd.concat([investment, gender], axis=1, ignore_index=True)
temp_df.columns = ["Investment", "Women Invester"]
temp_df = temp_df.groupby("Investment", as_index="Investment").sum()

raw_dataset_df = pd.read_csv(RAW_DATASET_LOC, low_memory=False, skiprows=[0])
raw_dataset_df = raw_dataset_df.iloc[:, :232]
raw_dataset_df.head()
raw_dataset_df.rename(columns={'Duration (in seconds)': 'Response Time',
'What is your age (# years)?': 'Age',
'What is your gender? - Selected Choice': 'Gender',
'In which country do you currently reside?': 'Country',
'What is the highest level of formal education that you have attained or plan to attain within the next 2 years?': 'Education',
```

```
'Select the title most similar to your current role (or most recent title if retired): - Selected Choice': 'Job Title',

'For how many years have you been writing code and/or programming?': 'Experience',

'Approximately how much money have you (or your team) spent on machine learning and/or cloud computing services at home (or at work) in the past 5 years (approximate $USD)?': 'Investment',

'What is your current yearly compensation (approximate $USD)?': 'Compensation'}, inplace=True, errors='raise')

raw_dataset_df
```

```
filerted_columns = ['Response Time', 'Age', 'Gender', 'Country', 'Education', 'Job Title', 'Experience', 'Investment', 'Compensation']

df = raw_dataset_df.filter(filerted_columns)

df.to_csv("Obj2_Dataset/PS3.csv", index=False)
```

```
filerted_columns = ['Response Time', 'Age', 'Gender', 'Country', 'Education', 'Job Title']

df = raw_dataset_df.filter(filerted_columns)

df.to_csv("Obj2_Dataset/PS2.csv", index=False)
```