Producing Visualization using Income Census Dataset

SUBMITTED BY:

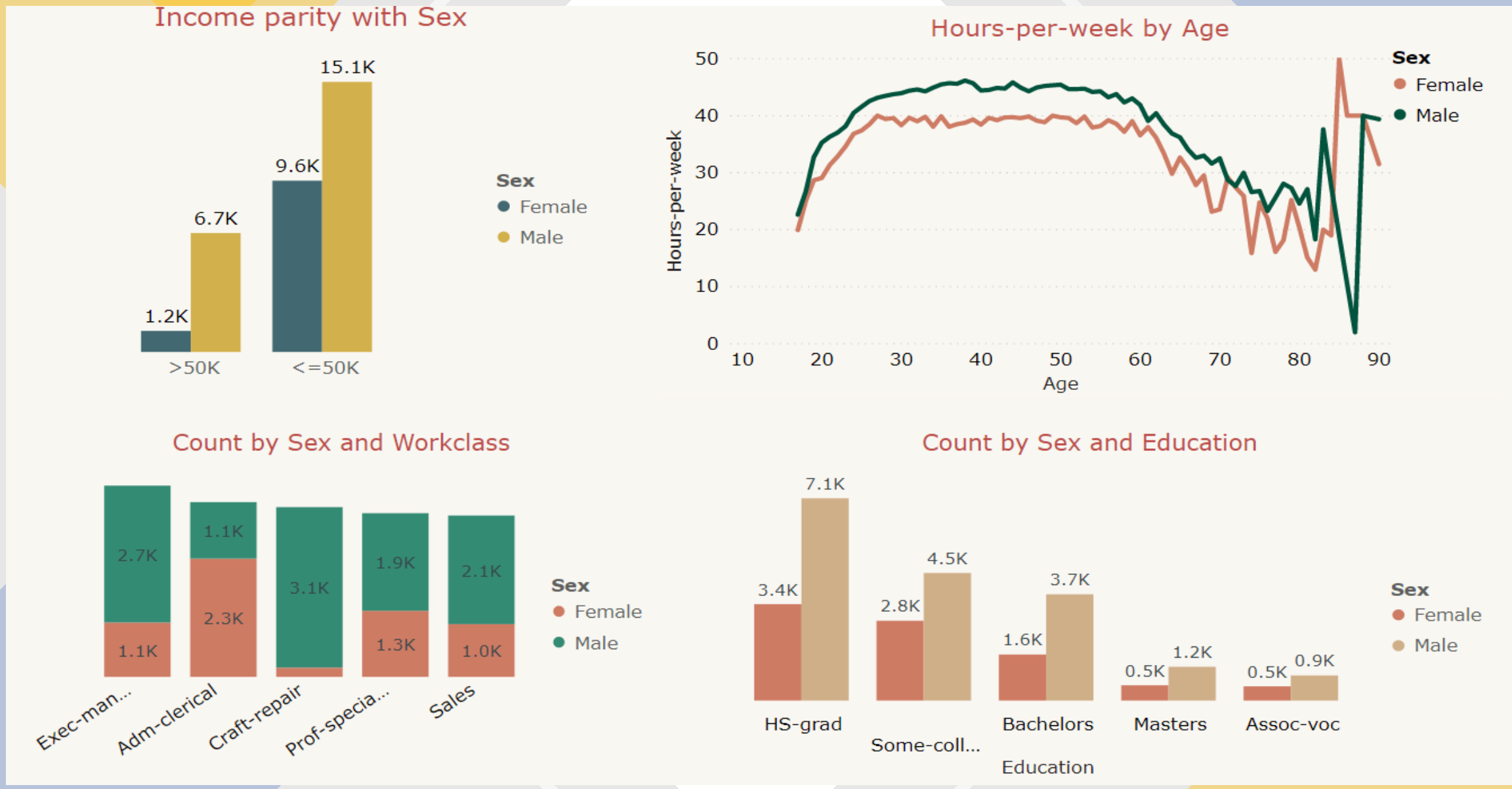
- **GARVIT MEHTA**
- **100809489**

Table of Contents

Questions we are trying to answer using the Income Census Dataset

- **Who is earning more… Male or Female ?**
- **Why are they earning more?**
- **Who is working for a greater number of hours ?**
- **What are the top earning Occupation ?**
- **Female Representation in these Occupations ?**

Visualization of the Dashboard



Data Description

UCI Machine Learning Repository

Data Set Information:

Extraction was done by Barry Becker from the 1994 Census database. A set of reasonably clean records was extracted using the following conditions: ((AAGE>16) && (AGI>100) && (AFNLWGT>1)&& (HRSWK>0))

Prediction task is to determine whether a person makes over 50K a year.

Attribute Information:

Listing of attributes:

>50K, <=50K.

age: continuous.
workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
fnlwgt: continuous.
education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
education-num: continuous.
marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
sex: Female, Male.
capital-gain: continuous.
capital-loss: continuous.
hours-per-week: continuous.
native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.

The Census Income dataset consists of 16 columns and 16384 Rows

Data Cleaning Part 1…

```
In [28]:
import pandas as pd
import numpy as np
```

```
In [29]:
#Load data set
df= pd.read_csv(r'C:\Users\garvi\OneDrive\Documents\Durham college\College Logo\income_evaluation - Copy.csv')
df.head(30)
```

Out[29]:

	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	income
0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
3	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
4	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K
5	37	Private	284582	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	40	United-States	<=50K
6	49	Private	160187	9th	5	Married-spouse-absent	Other-service	Not-in-family	Black	Female	0	0	16	Jamaica	<=50K
7	52	Self-emp-not-inc	209642	HS-grad	9	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	45	United-States	>50K
8	31	Private	45781	Masters	14	Never-	Prof-	Not-in-family	White	Female	14084	0	50	United-	>50K

Dropping the columns

…Data Cleaning Part 1

Dropped the columns

In [30]:
```
#Dropping columns from the dataframe
df= df.drop([' education-num', ' capital-gain', ' capital-loss',' fnlwgt'], axis=1)
df.head(30)
```

Out[30]:

	age	workclass	education	marital-status	occupation	relationship	race	sex	hours-per-week	native-country	income
0	39	State-gov	Bachelors	Never-married	Adm-clerical	Not-in-family	White	Male	40	United-States	<=50K
1	50	Self-emp-not-inc	Bachelors	Married-civ-spouse	Exec-managerial	Husband	White	Male	13	United-States	<=50K
2	38	Private	HS-grad	Divorced	Handlers-cleaners	Not-in-family	White	Male	40	United-States	<=50K
3	53	Private	11th	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	40	United-States	<=50K
4	28	Private	Bachelors	Married-civ-spouse	Prof-specialty	Wife	Black	Female	40	Cuba	<=50K
5	37	Private	Masters	Married-civ-spouse	Exec-managerial	Wife	White	Female	40	United-States	<=50K
6	49	Private	9th	Married-spouse-absent	Other-service	Not-in-family	Black	Female	16	Jamaica	<=50K
7	52	Self-emp-not-inc	HS-grad	Married-civ-spouse	Exec-managerial	Husband	White	Male	45	United-States	>50K
8	31	Private	Masters	Never-married	Prof-specialty	Not-in-family	White	Female	50	United-States	>50K
9	42	Private	Bachelors	Married-civ-spouse	Exec-managerial	Husband	White	Male	40	United-States	>50K
10	37	Private	Some-college	Married-civ-spouse	Exec-managerial	Husband	Black	Male	80	United-States	>50K
11	30	State-gov	Bachelors	Married-civ-spouse	Prof-specialty	Husband	Asian-Pac-Islander	Male	40	India	>50K
12	23	Private	Bachelors	Never-married	Adm-clerical	Own-child	White	Female	30	United-States	<=50K

Data Cleaning Part 2

27	54	?	Some-college	Married-civ-spouse	?	Husband	Asian-Pac-Islander	Male	60	South	>50K
28	39	Private	HS-grad	Divorced	Exec-managerial	Not-in-family	White	Male	80	United-States	<=50K
29	49	Private	HS-grad	Married-civ-spouse	Craft-repair	Husband	White	Male	40	United-States	<=50K

In [31]:
```
df.isnull().sum()
```

Out[31]:
```
age               0
 workclass        0
 education        0
 marital-status   0
 occupation       0
 relationship     0
 race             0
 sex              0
 hours-per-week   0
 native-country   0
 income           0
dtype: int64
```

In [32]:
```
#replacing '?' with Unknown:
df[' workclass'].replace(' ?', ' Unknown',inplace=True)
df[' occupation'].replace(' ?', ' Unknown',inplace=True)
df[' native-country'].replace(' ?', ' Unknown',inplace=True)
```

In [26]:
```
df.head(30)
```

Out[26]:

	age	workclass	education	marital-status	occupation	relationship	race	sex	hours-per-week	native-country	income
0	39	State-gov	Bachelors	Never-married	Adm-clerical	Not-in-family	White	Male	40	United-States	<=50K

Checking for the Null values and replacing '?' with 'Unknown'

…Data Cleaning Part 2

Replaced the '?' with 'Unknown'

In [32]:
```
#replacing '?' with Unknown:
df[' workclass'].replace(' ?', ' Unknown',inplace=True)
df[' occupation'].replace(' ?', ' Unknown',inplace=True)
df[' native-country'].replace(' ?', ' Unknown',inplace=True)
```

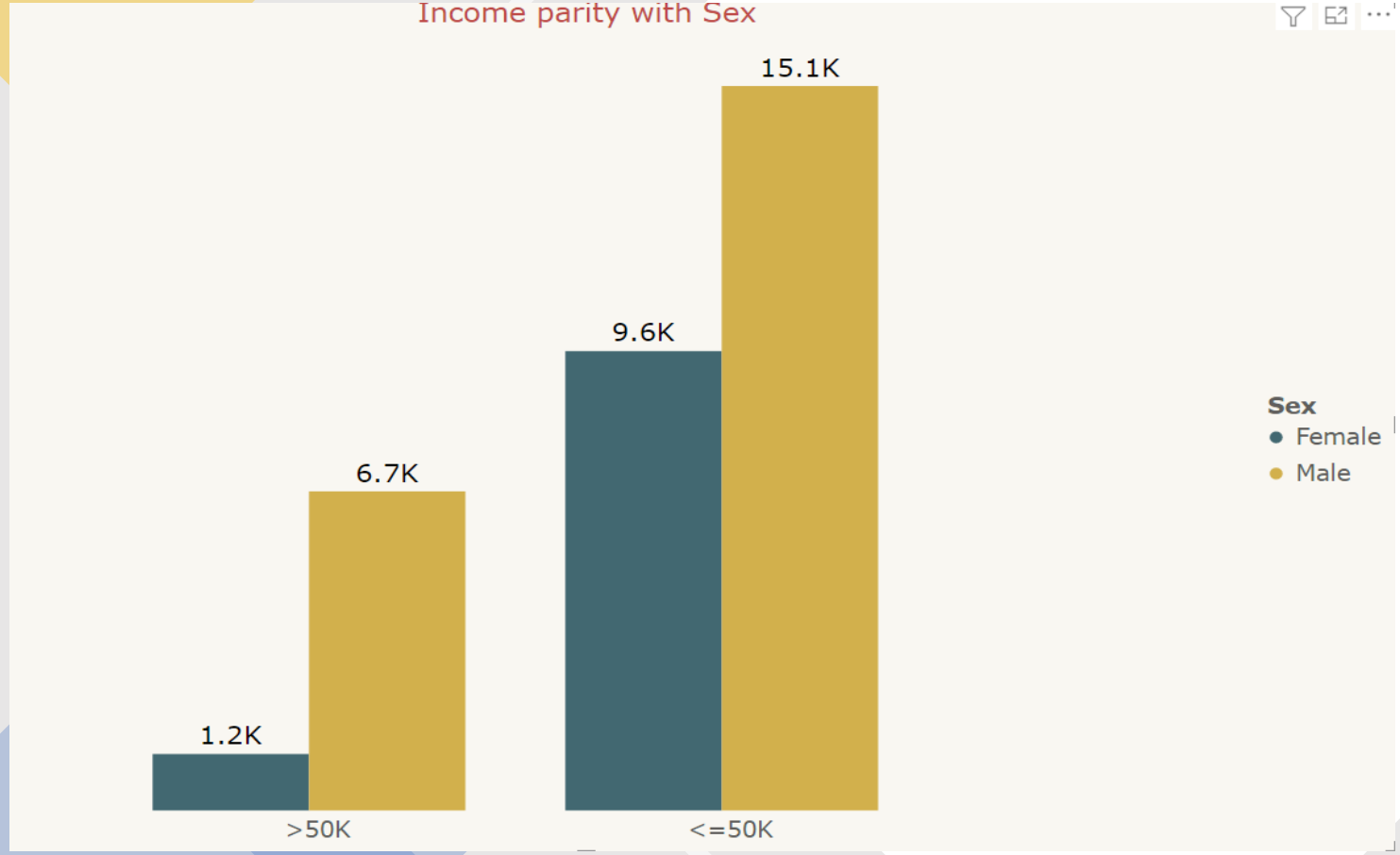
In [26]:
```
df.head(30)
```

					inspct						
18	38	Private	11th	Married-civ-spouse	Sales	Husband	White	Male	50	United-States	<=50K
19	43	Self-emp-not-inc	Masters	Divorced	Exec-managerial	Unmarried	White	Female	45	United-States	>50K
20	40	Private	Doctorate	Married-civ-spouse	Prof-specialty	Husband	White	Male	60	United-States	>50K
21	54	Private	HS-grad	Separated	Other-service	Unmarried	Black	Female	20	United-States	<=50K
22	35	Federal-gov	9th	Married-civ-spouse	Farming-fishing	Husband	Black	Male	40	United-States	<=50K
23	43	Private	11th	Married-civ-spouse	Transport-moving	Husband	White	Male	40	United-States	<=50K
24	59	Private	HS-grad	Divorced	Tech-support	Unmarried	White	Female	40	United-States	<=50K
25	56	Local-gov	Bachelors	Married-civ-spouse	Tech-support	Husband	White	Male	40	United-States	>50K
26	19	Private	HS-grad	Never-married	Craft-repair	Own-child	White	Male	40	United-States	<=50K
27	54	Unknown	Some-college	Married-civ-spouse	Unknown	Husband	Asian-Pac-Islander	Male	60	South	>50K
28	39	Private	HS-grad	Divorced	Exec-managerial	Not-in-family	White	Male	80	United-States	<=50K

In [39]:
```
#Saving our cleaned file as CSV
df.to_csv(r'C:\Users\garvi\OneDrive\Documents\Durham college\College Logo\income_evaluation - Cleaned_Copy.csv', index=False)
```

In [ ]:

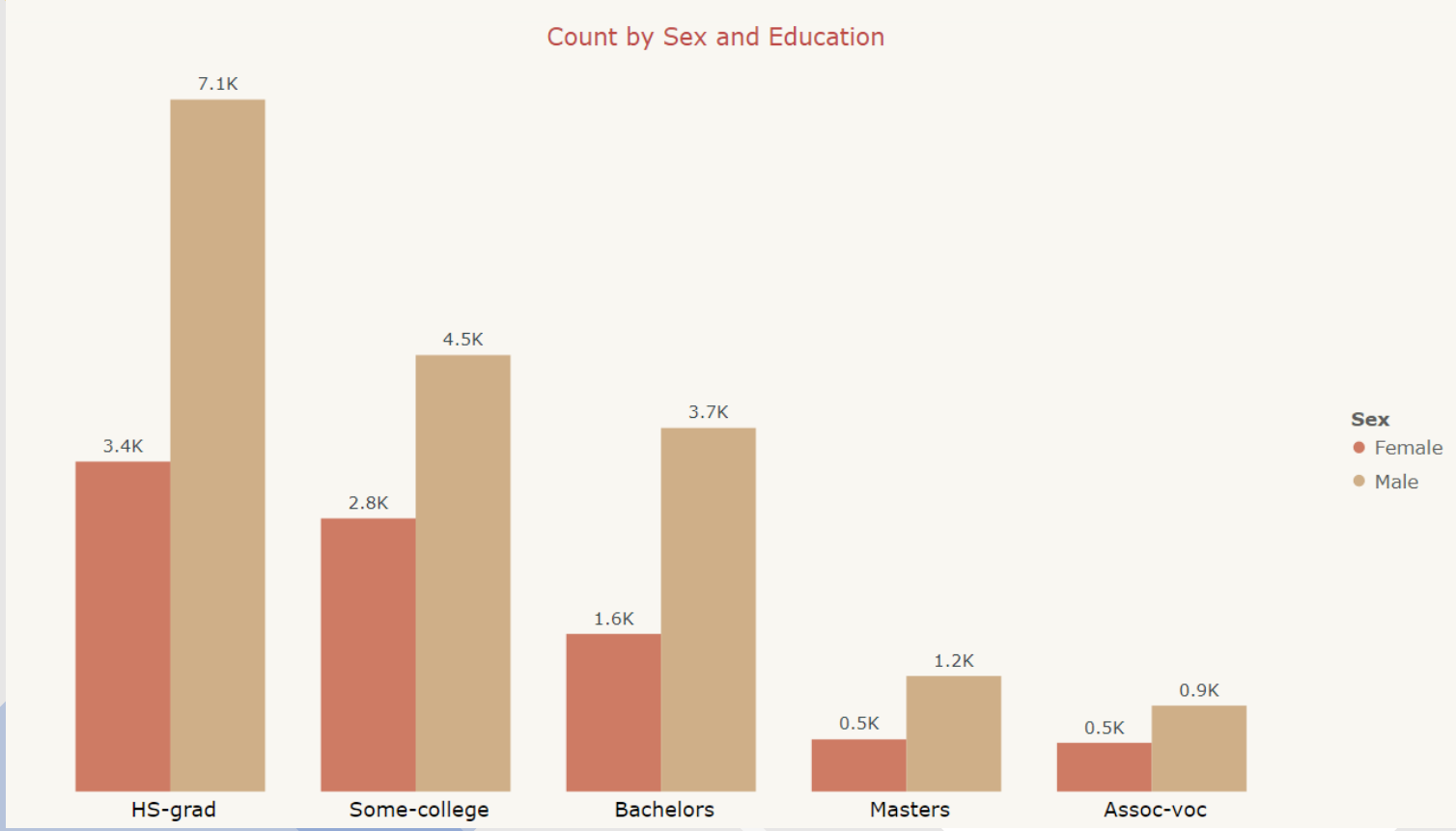
Who is earning more….?



Inferences from the chart

- A clustered column chart is used for this visualization. The **Income category is at X axis** and the **count** of the **Males** and **Females** on **Y axis**.
- It can be inferred from the chart that number of **Males earning more than $50k** is almost **five times** the number of **Females**.
- **Men** are **more likely** to **earn more** than **$50K**

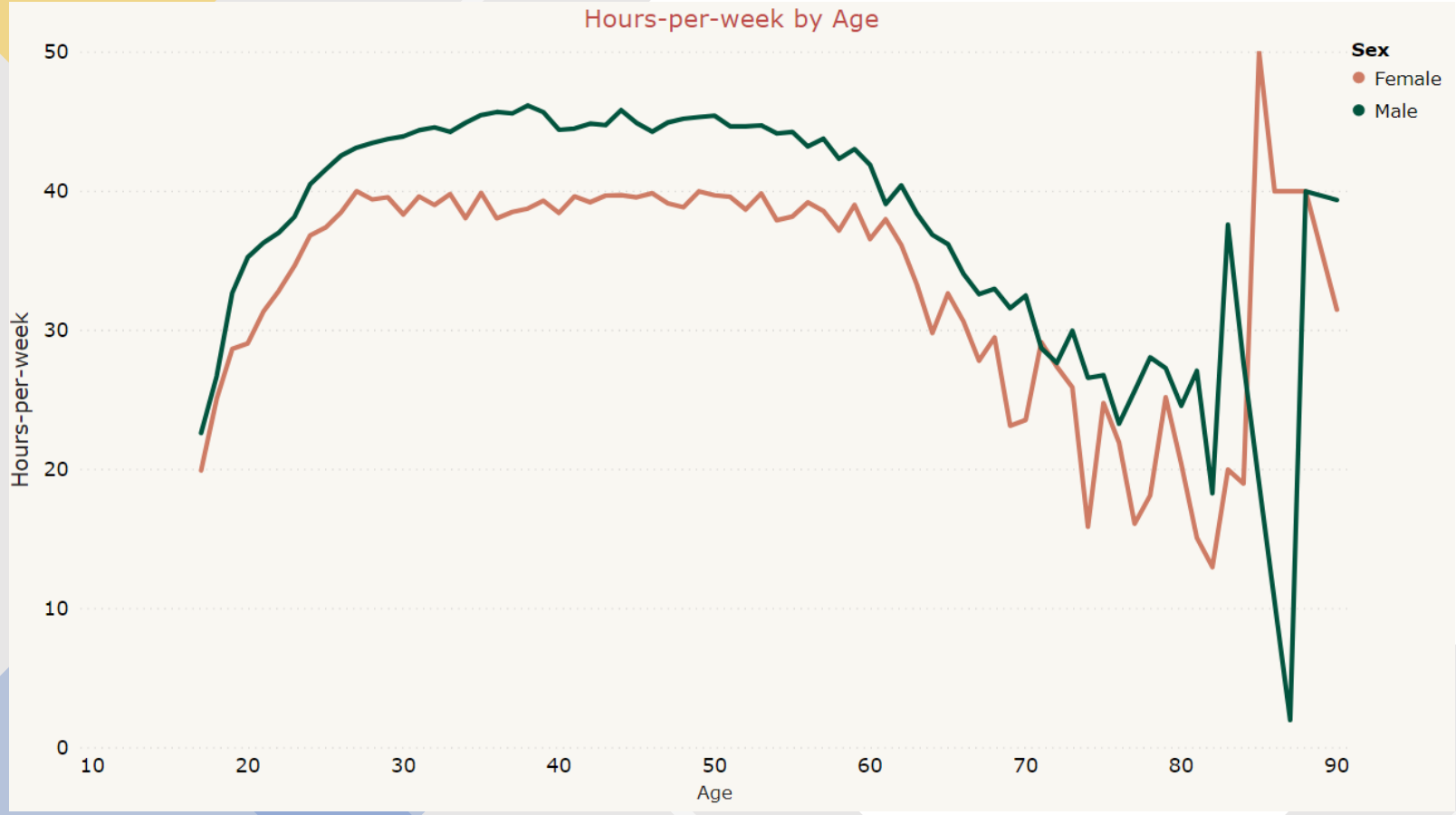
Why are they earning more…. ?



Inferences from the chart

- A Stacked Column Chart has been used to show the **count of people (Y axis)** based on the level of **Education** (X axis) attained by them.
- The visual shows that more Men are likely to attain a certain level of education as compared to their female counterpart.
- No surprises here as to why Men are earning more than females. Since men are more educated than females they are better paid off for their job and are earning handsomely.

Who is working for a greater number of hours….?

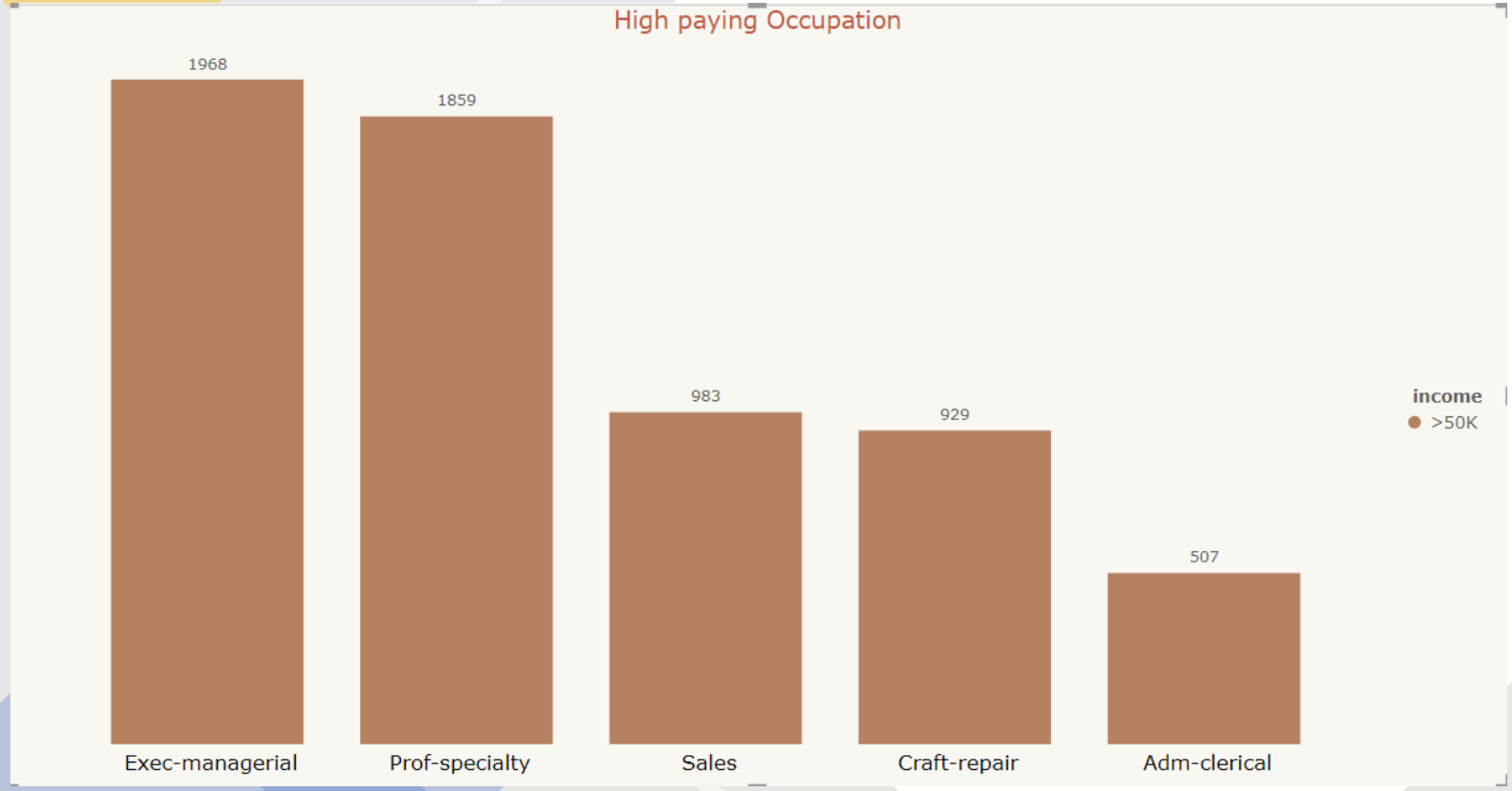


Inferences from the chart

- A line chart has been used to establish the relation between **Hours per week (Y axis)** spent by an individual **(Sex as legend)** and their **Age (X axis).**

- We can see that **Males** put in **more no. of hours** than Females for almost all Ages. This trend continues **till age 85**, after which **females have worked more**. In the age range from **85-88, females** have **outperformed** the males, thereafter which males again worked for more no of hours. Overall, we can say **Males have put more number of hours working than Females.**

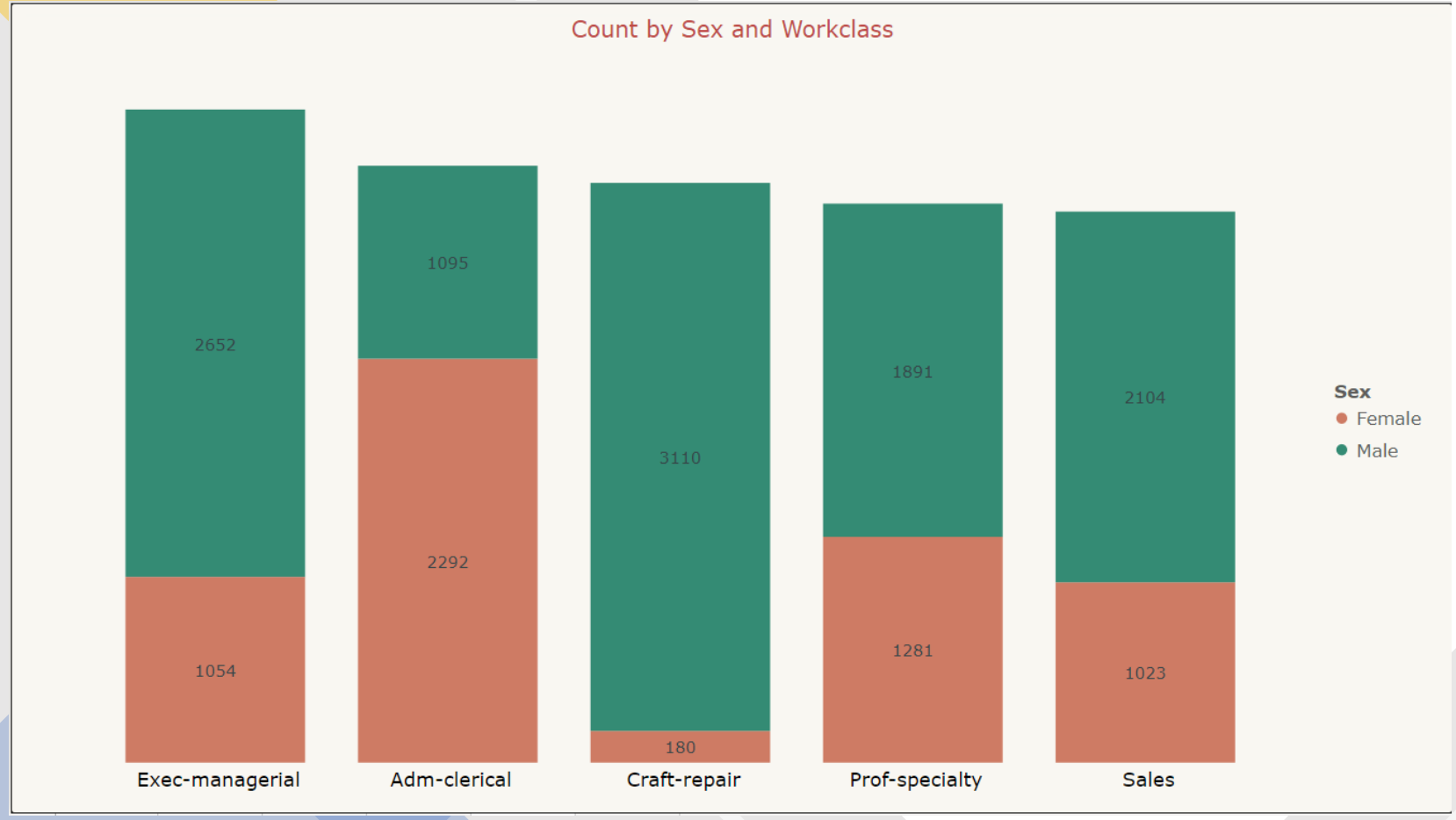
What are the high paying Occupations….?



Inferences from the chart

- A Clustered Column Chart is used for this visualization. Type of **Occupation** is in **(X axis)** and **Count of people** under that occupation is in **(Y axis).**
- This chart presents the **number of people employed in occupations which are paying more than $50k.**
- It can be seen from the chart that **Exec-managerial** is the **highest paying** occupation with **1968** professionals working in it, followed by **Prof-specialty** with **1859** and last is **Adm-clerical** with **507** professionals employed in the same occupation.

Female representation in these Occupations….?



Inferences from the chart

- A Stacked column chart is used for this visualization. The Occupation is in X axis and Count of people is in Y axis. It is further filtered by the gender of these people.
- The chart represents the **Highest paying occupation** with **Female representation**. The **number of females** in any occupation (highest paying occupation) are **less** compared to their male counter parts **except the Adm-clerical job** where females have out played males.
- Since **less females take up** the **highest paying** occupation, we can easily understand why **they are getting paid less** compared to males.

Conclusions

According to the Income Census data it can be concluded that **Males earn more than Females**, this can be attributed to the following facts:

- Males are **more educated** as compared to their counter parts for any level of education.
- They can devote **more hours at work** as compared to women.
- Males have **more opportunities** to work in highest paying Professions, whereas, Females have less representation in these Occupations.

Video Link

Please find the link below to my YouTube channel for the Interactive Video presentation:

https://youtu.be/V9tgT2PcVfY

References

- The Census Income dataset has been taken from: UCI machine learning repository.

- *Data Link: https://archive.ics.uci.edu/ml/datasets/census+income*

Thank You