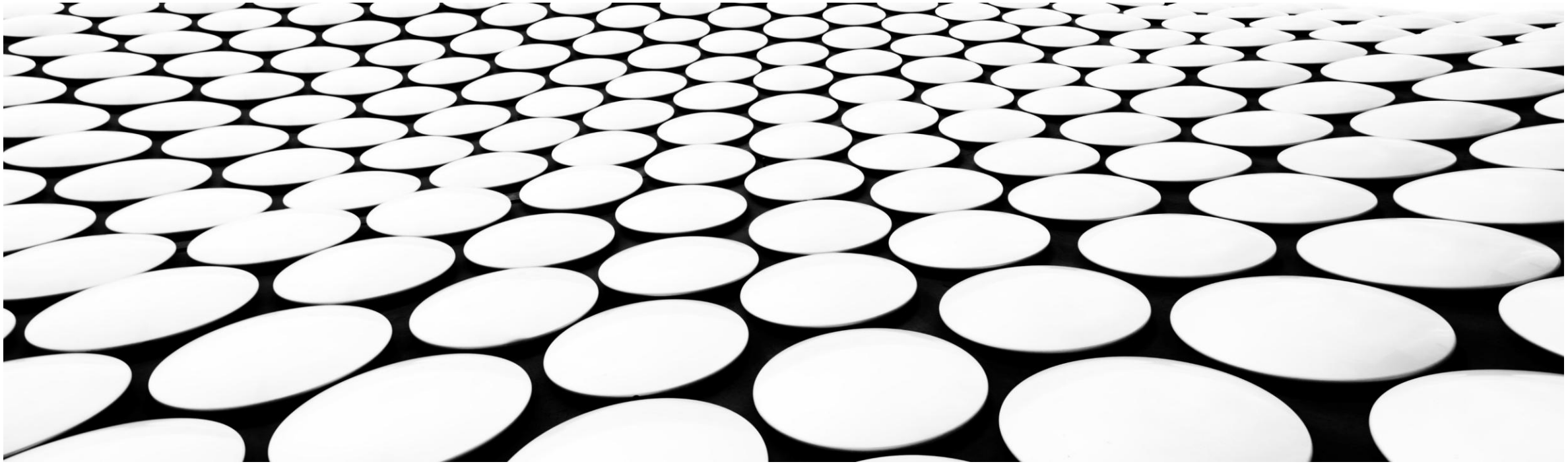
EXPLORATORY DATA ANALYSIS

Story behind the data: Census dataset

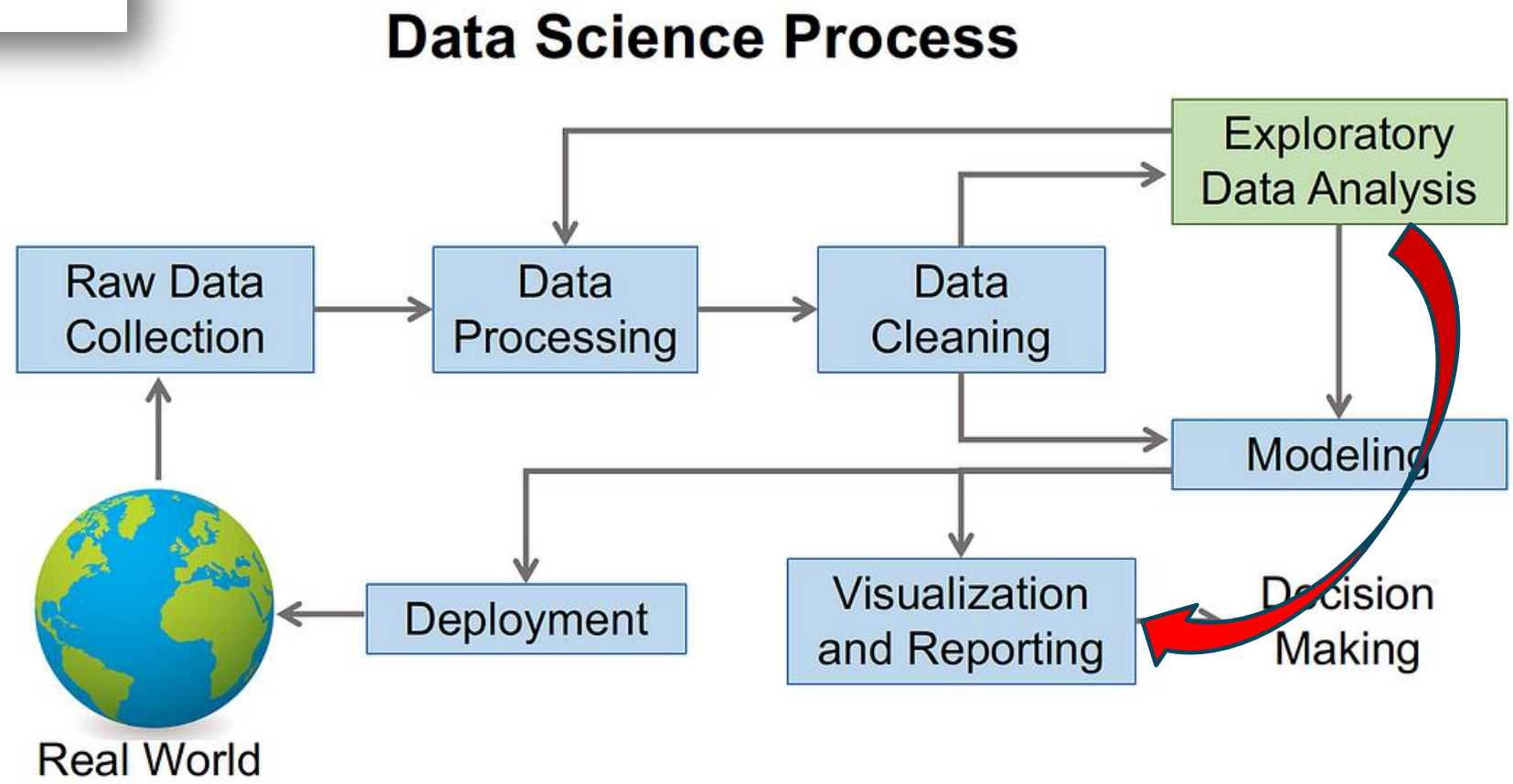


EXPLORATORY DATA ANALYSIS

- Case Study 2 – Census data
- Assignment 1 Examples

Data Science Process: A Comprehensive Guide

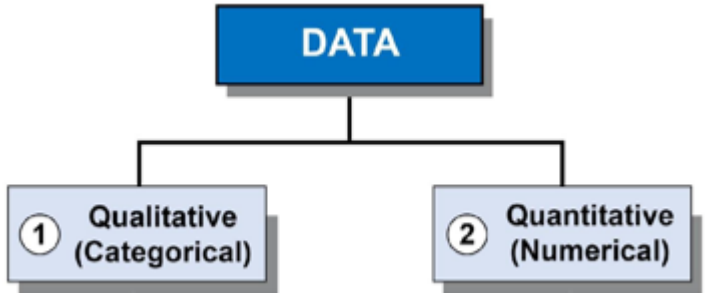
Abhijit · Follow
6 min read · Jan 15, 2024



Source

Univariate Analysis

- Numerical feature
- Categorical feature



Bivariate Analysis

- Numerical/Categorical
- Categorical/Categorical
- Categorical/Numerical
- Numerical/Numerical

We didn't have time to explore those…

Case Study 2
Census Dataset

UC Irvine Machine Learning Repository

Adult

Donated on 4/30/1996

Predict whether annual income of an individual exceeds $50K/yr based on census data. Also known as "Census Income" dataset.

Dataset Characteristics	Subject Area	Associated Tasks
Multivariate	Social Science	Classification
Feature Type	**# Instances**	**# Features**
Categorical, Integer	48842	14

Explore the dataset

https://archive.ics.uci.edu/dataset/2/adult

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32561 entries, 0 to 32560
Data columns (total 15 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   age             32561 non-null  int64
 1   work-class      30725 non-null  object
 2   fnlwgt          32561 non-null  int64
 3   education       32561 non-null  object
 4   education-num   32561 non-null  int64
 5   marital-status  32561 non-null  object
 6   occupation      30718 non-null  object
 7   relationship    32561 non-null  object
 8   race            32561 non-null  object
 9   sex             32561 non-null  object
 10  capital-gain    32561 non-null  int64
 11  capital-loss    32561 non-null  int64
 12  hours-per-week  32561 non-null  int64
 13  native-country  31978 non-null  object
 14  income          32561 non-null  object
```

Look at a bit of data...

	age	work-class	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

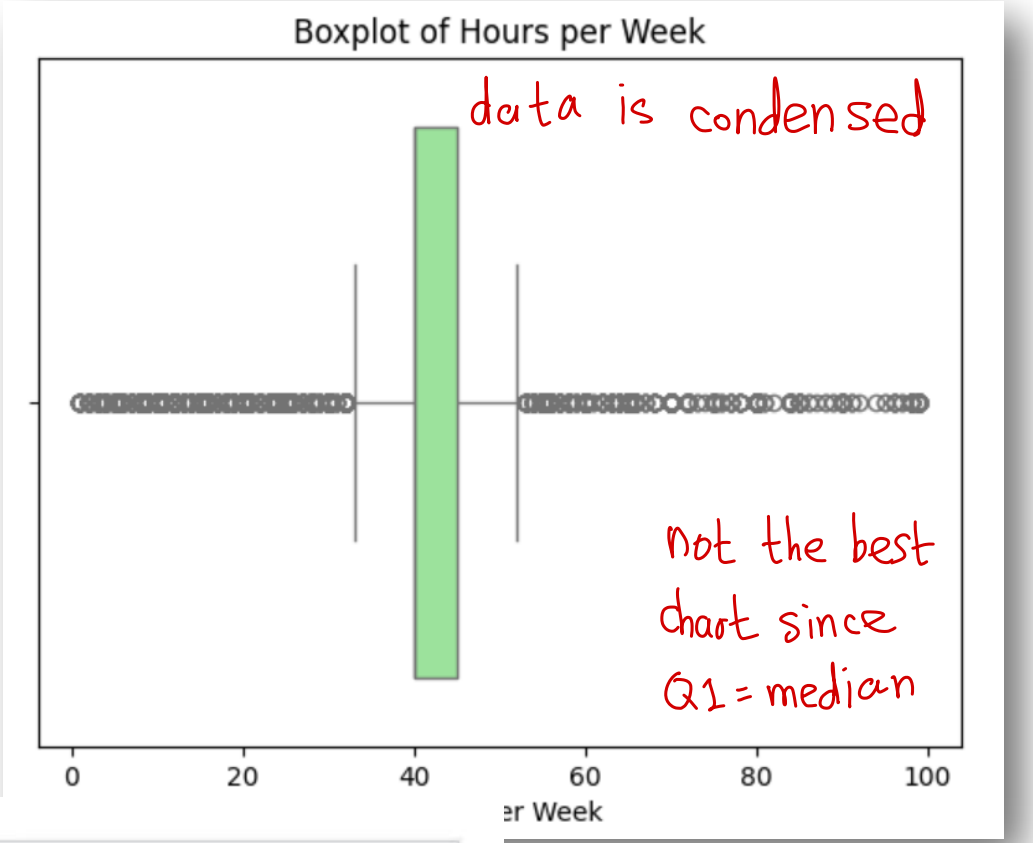
Surveyed

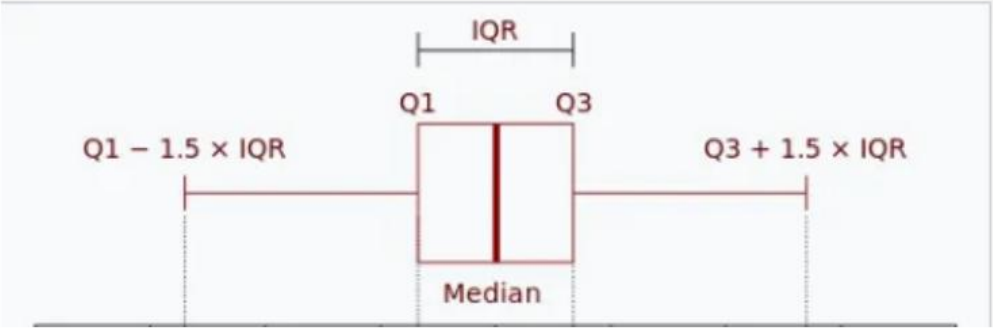
what is fnlwgt in the census dataset from uci

> `fnlwgt` is a sampling weight indicating how many people in the U.S. population a given individual represents.

Univariate analysis Numerical Data

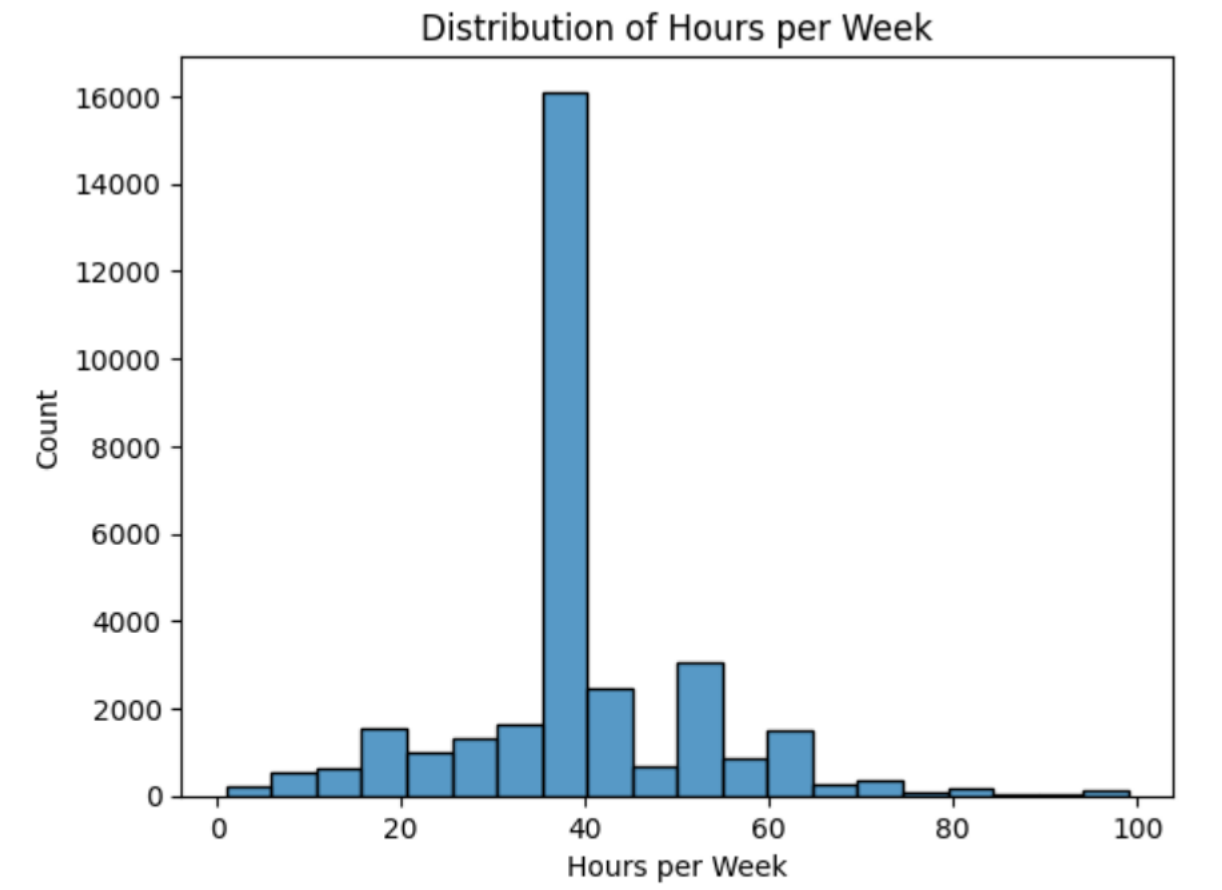
	hours-per-week
count	32561.000000
mean	40.437456
std	12.347429
min	1.000000
25%	40.000000
50%	40.000000
75%	45.000000
max	99.000000

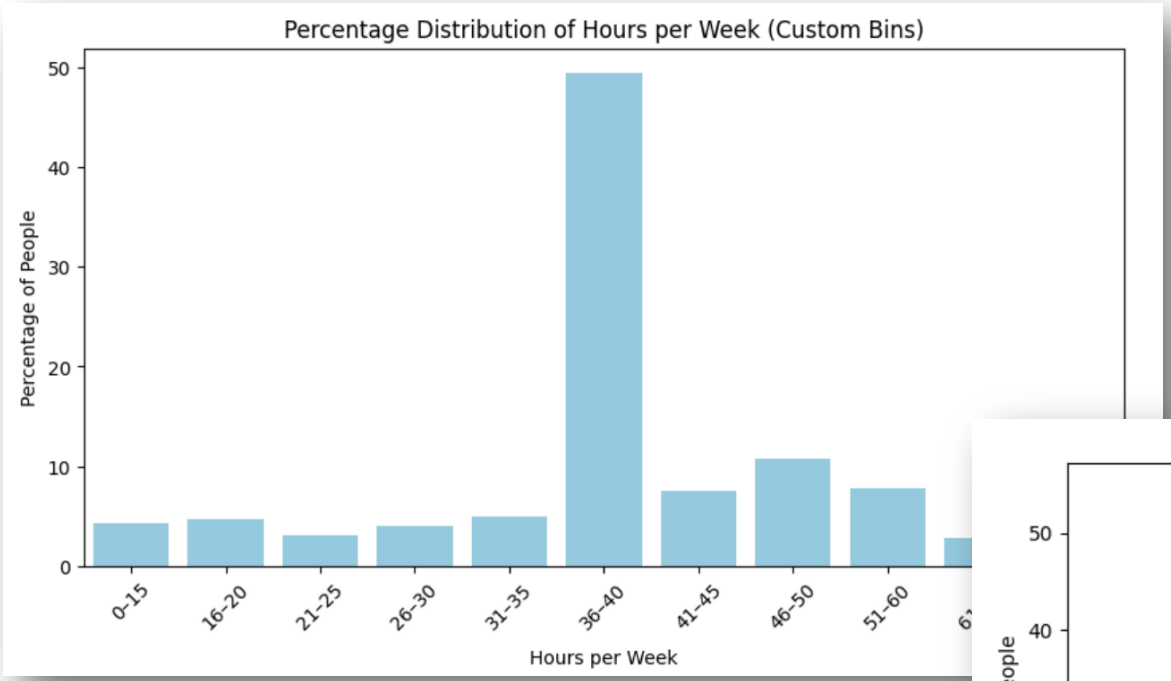




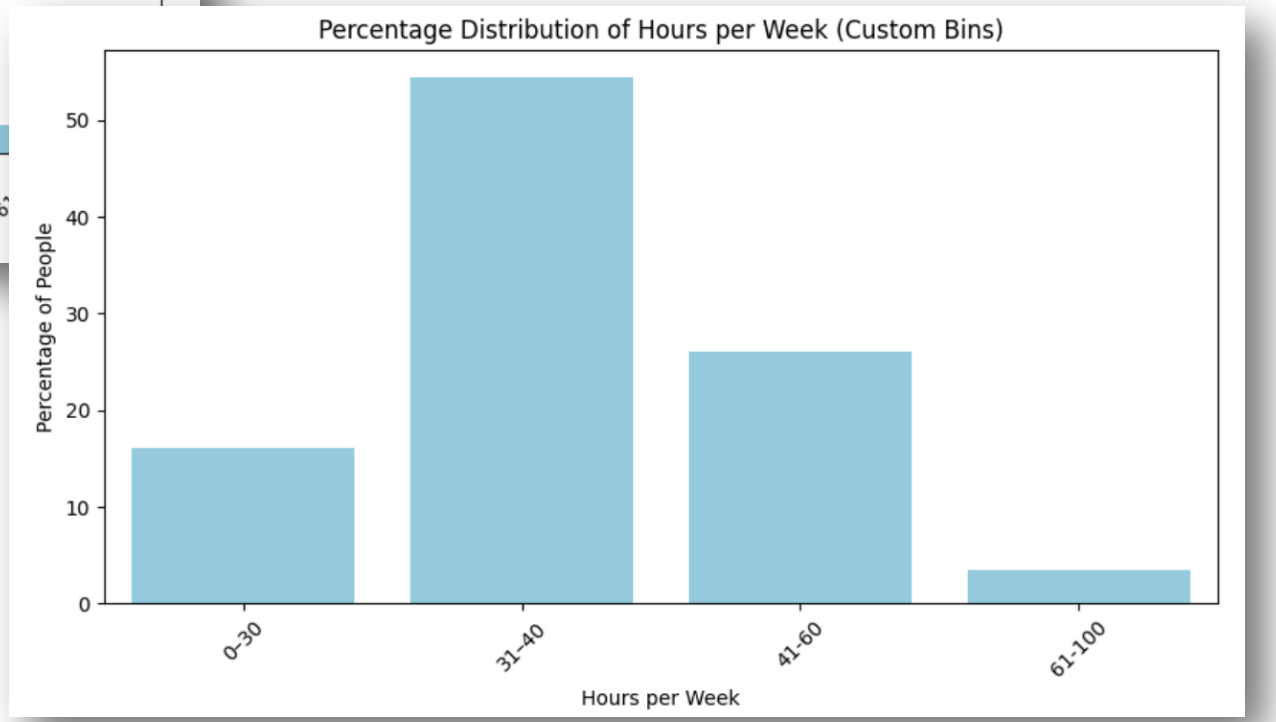
```
census['hours-per-week'].describe()
```

	hours-per-week
count	32561.000000
mean	40.437456
std	12.347429
min	1.000000
25%	40.000000
50%	40.000000
75%	45.000000
max	99.000000

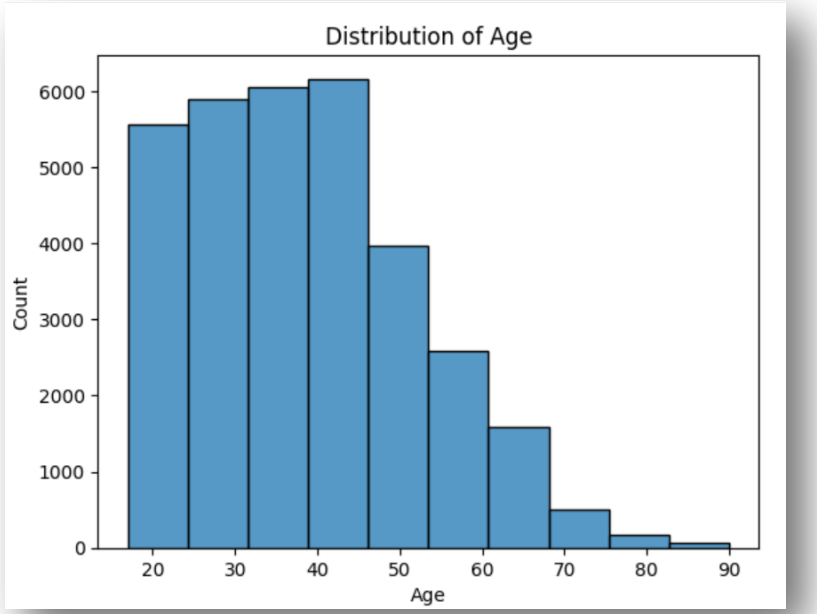


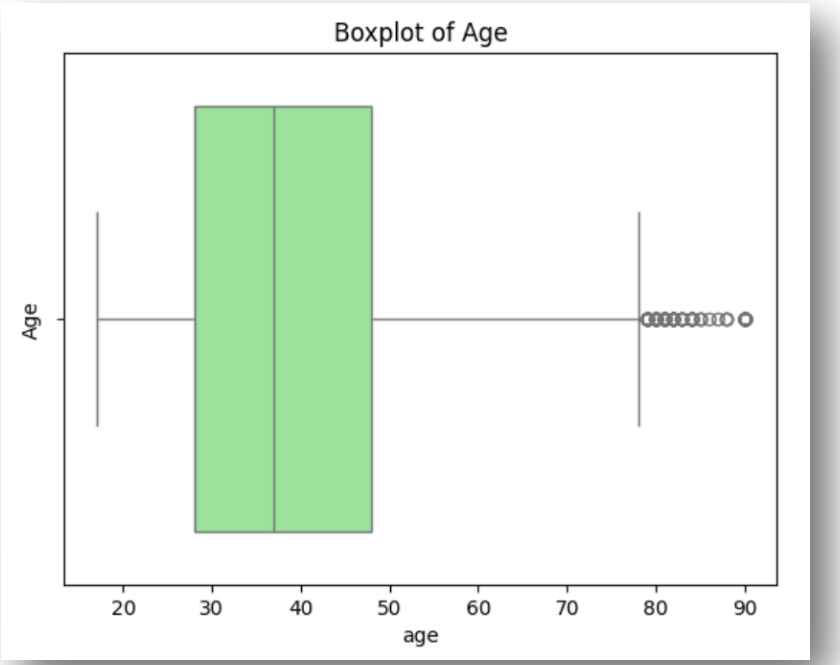


Binning introduces a type of biais...



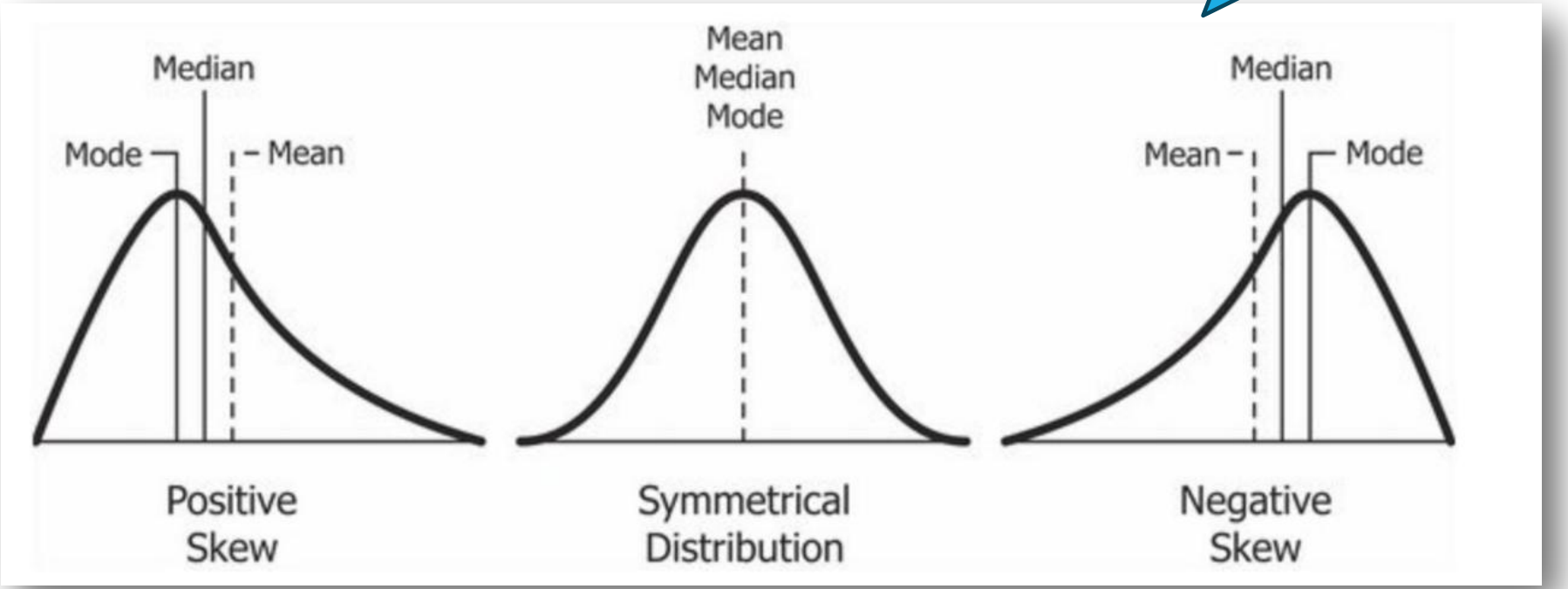
	age
count	32561.000000
mean	38.581647
std	13.640433
min	17.000000
25%	28.000000
50%	37.000000
75%	48.000000
max	90.000000





WIKIPEDIA
The Free Encyclopedia

Name refers to direction of the tail

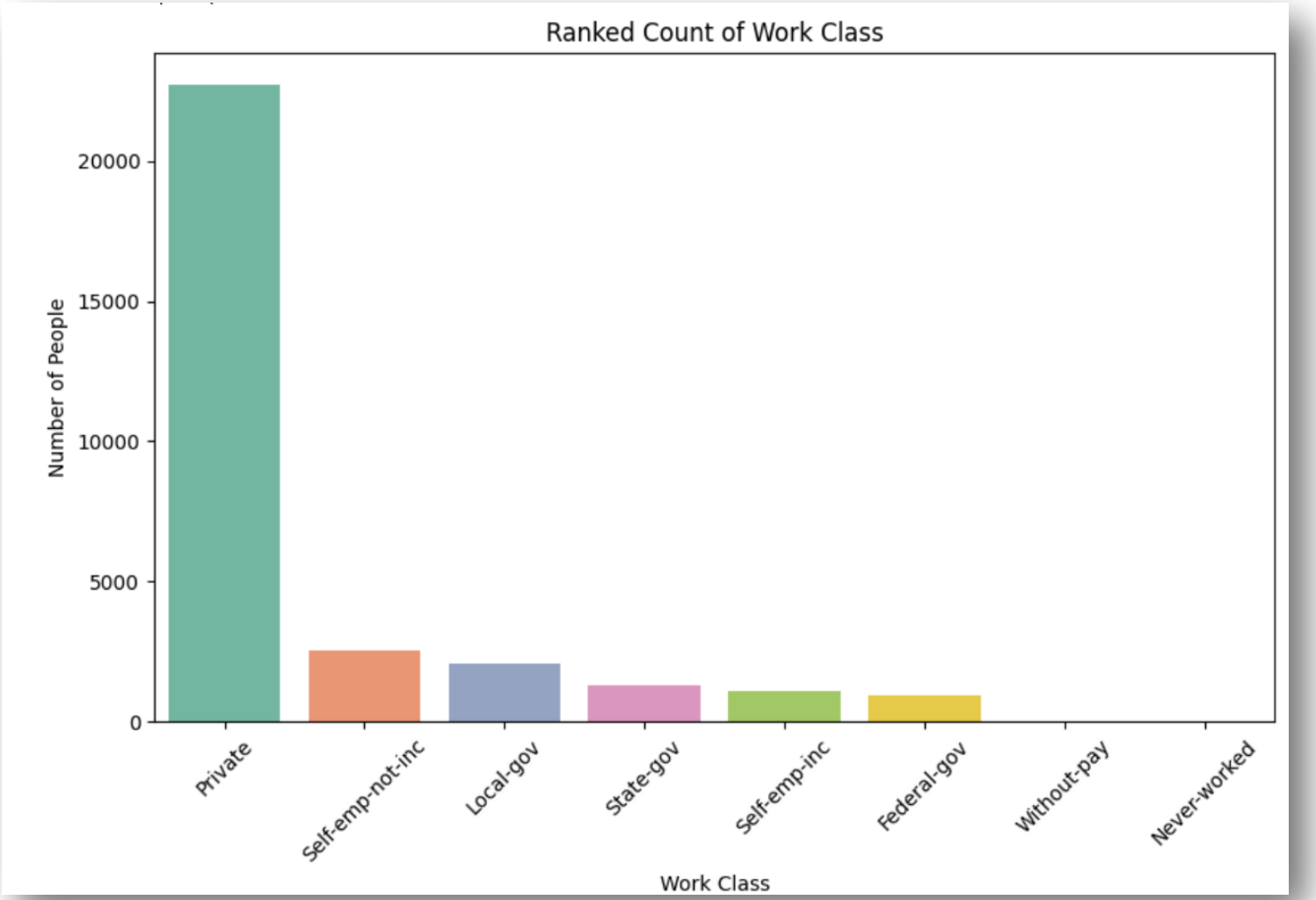


https://en.wikipedia.org/wiki/Skewness

Univariate analysis Categorical Data

```
census.value_counts("work-class")
```

work-class	count
Private	22696
Self-emp-not-inc	2541
Local-gov	2093
State-gov	1298
Self-emp-inc	1116
Federal-gov	960
Without-pay	14
Never-worked	7



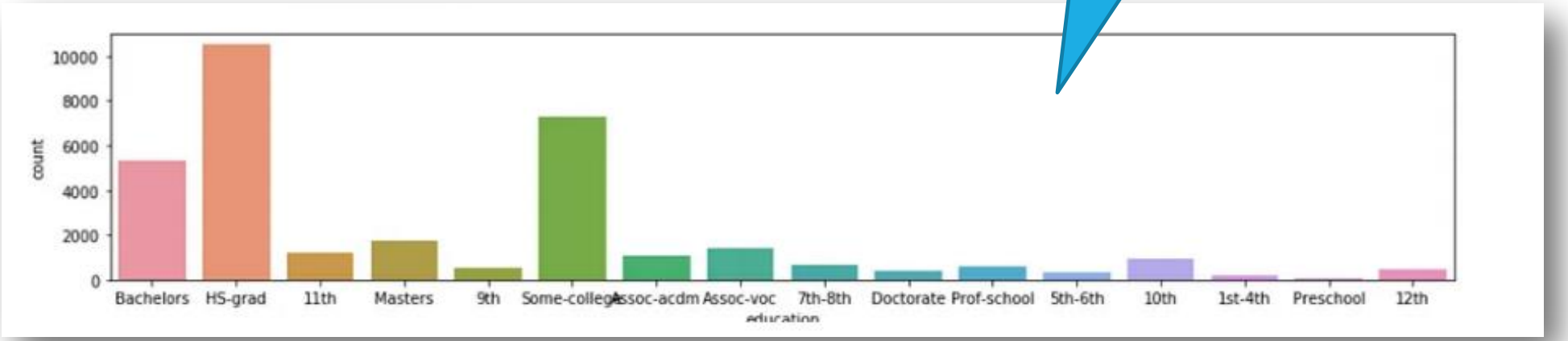
We don't know meaning of the variable names

Add a legend

```
census.value_counts("education")
```

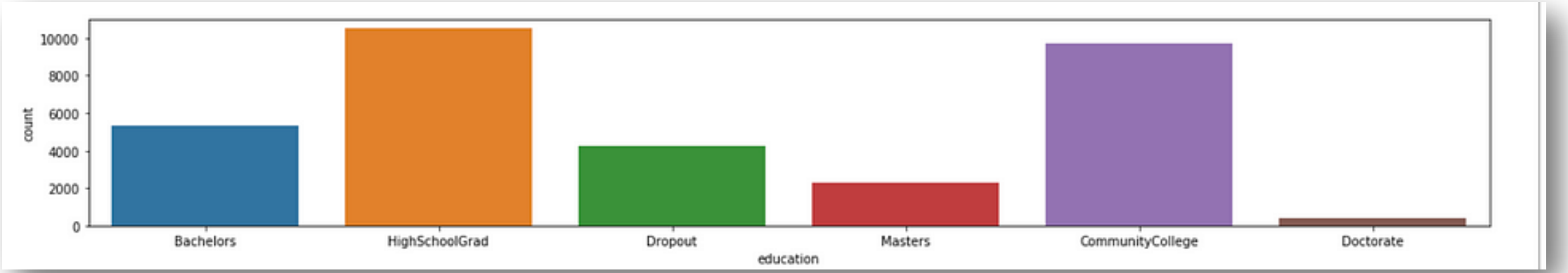
education	count
HS-grad	10501
Some-college	7291
Bachelors	5355
Masters	1723
Assoc-voc	1382
11th	1175
Assoc-acdm	1067
10th	933
7th-8th	646
Prof-school	576
9th	514
12th	433
Doctorate	413
5th-6th	333
1st-4th	168
Preschool	51

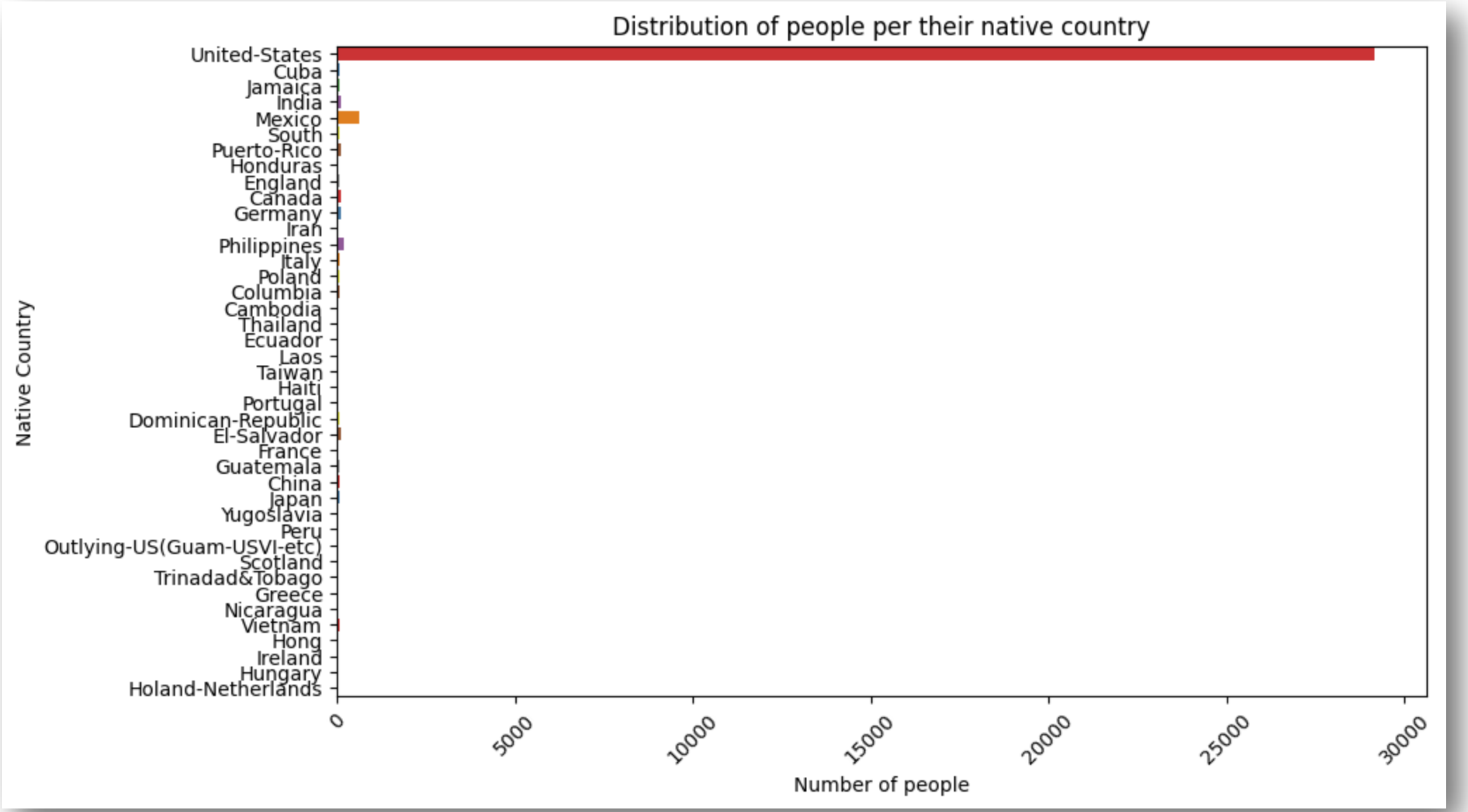
Too many possible values, hard to interpret

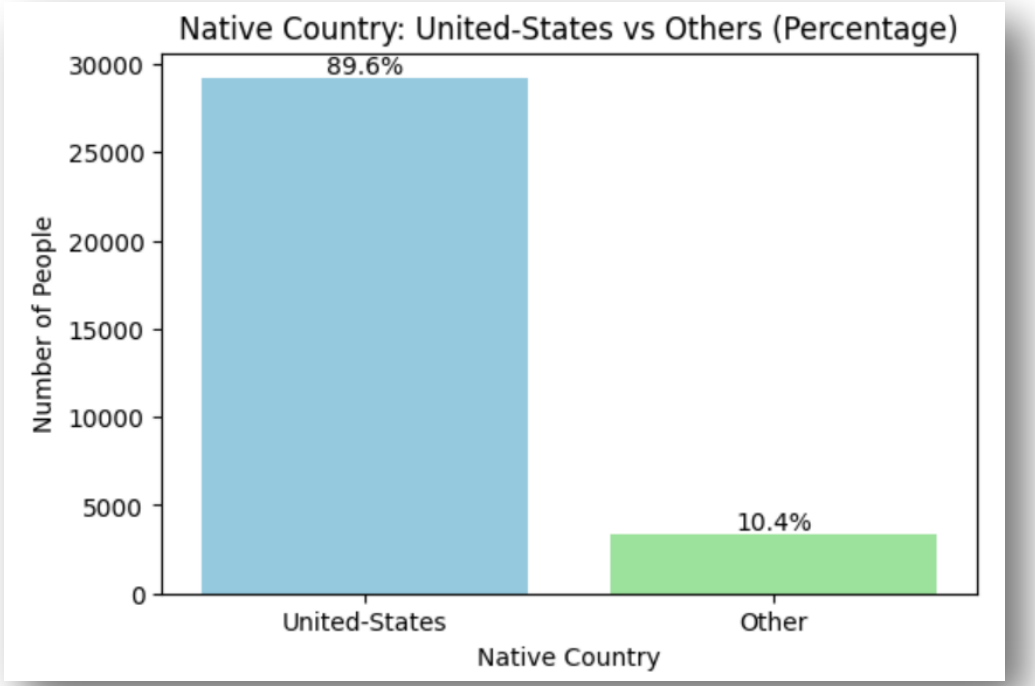


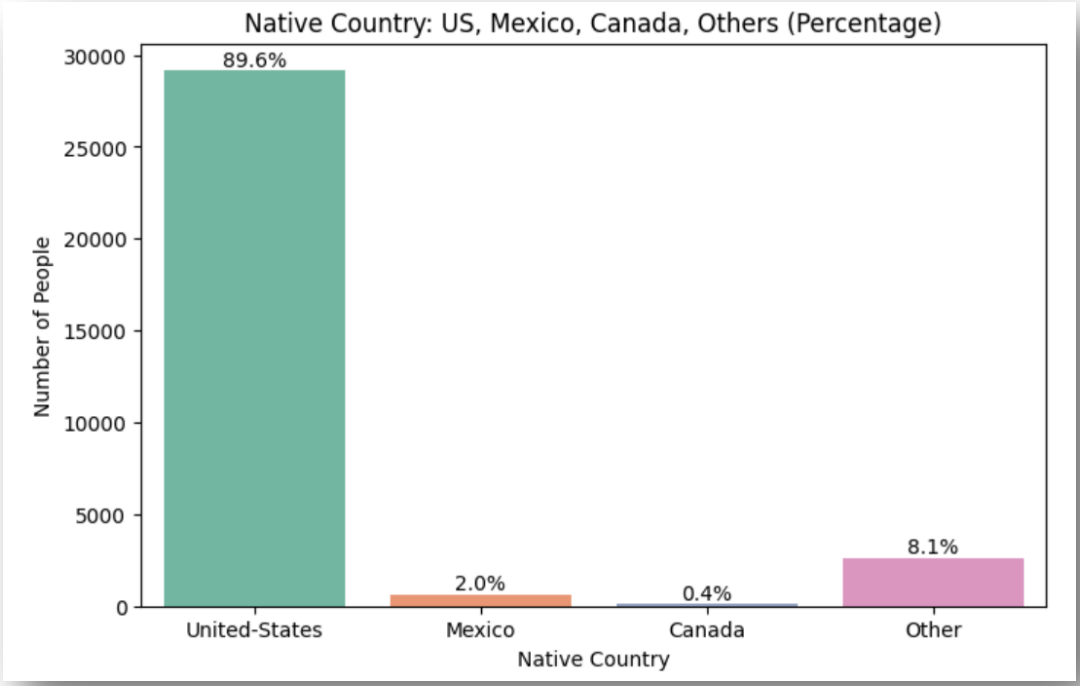
In [28]:
```
#Grouping education
data['education'].replace('Preschool', 'Dropout',inplace=True)
data['education'].replace('10th', 'Dropout',inplace=True)
data['education'].replace('11th', 'Dropout',inplace=True)
data['education'].replace('12th', 'Dropout',inplace=True)
data['education'].replace('1st-4th', 'Dropout',inplace=True)
data['education'].replace('5th-6th', 'Dropout',inplace=True)
data['education'].replace('7th-8th', 'Dropout',inplace=True)
data['education'].replace('9th', 'Dropout',inplace=True)
data['education'].replace('HS-Grad', 'HighSchoolGrad',inplace=True)
data['education'].replace('HS-grad', 'HighSchoolGrad',inplace=True)
data['education'].replace('Some-college', 'CommunityCollege',inplace=True)
data['education'].replace('Assoc-acdm', 'CommunityCollege',inplace=True)
data['education'].replace('Assoc-voc', 'CommunityCollege',inplace=True)
data['education'].replace('Prof-school', 'Masters',inplace=True)

fig = plt.figure(figsize=(20,3))
sns.countplot(x="education", data=data)
```

Domain knowledge necessary to group categories, do you agree with their groups?




Distribution of people per their native country
United-States
Cuba
Jamaica
India
Mexico
South
Puerto-Rico
Honduras
England
Canada
Germany
Iran
Philippines
Italy
Poland
Columbia
Cambodia
Thailand
Ecuador
Laos
Taiwan
Haiti
Portugal
Dominican-Republic
El-Salvador
France
Guatemala
China
Japan
Yugoslavia
Peru
Outlying-US(Guam-USVI-etc)
Scotland
Trinadad&Tobago
Greece
Nicaragua
Vietnam
Hong
Ireland
Hungary
Holand-Netherlands
Native Country
0
5000
10000
15000
20000
25000
30000
Number of people

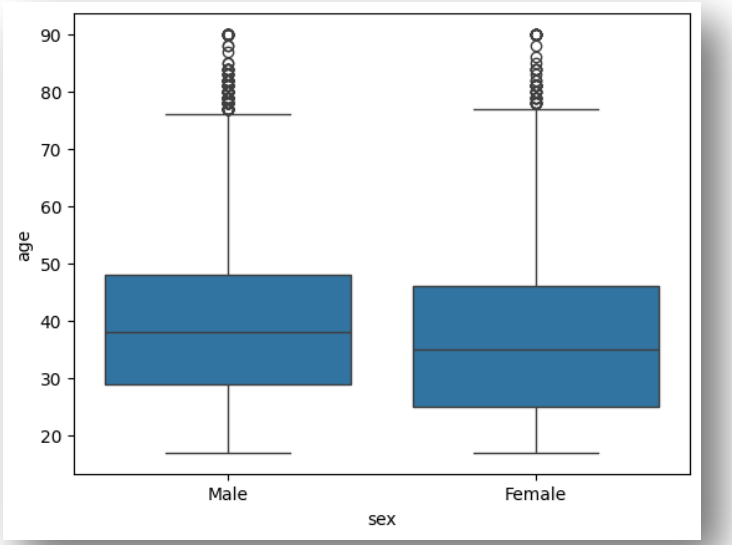




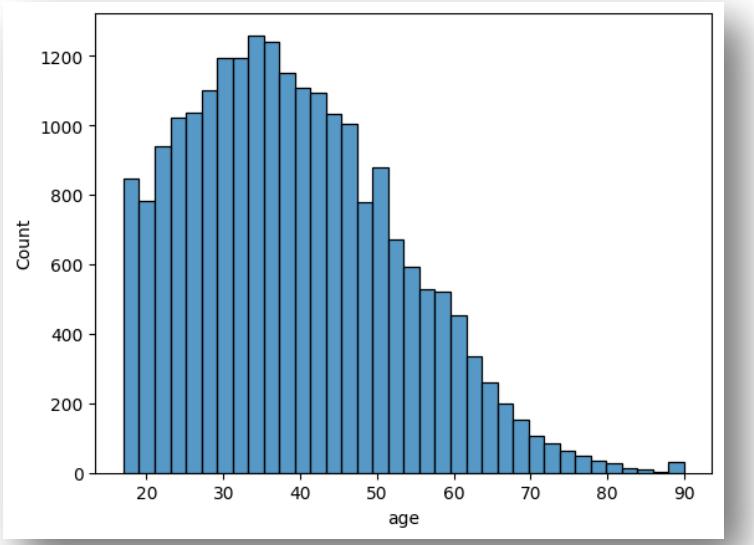
Grouping introduces a particular view

Can introduce bias

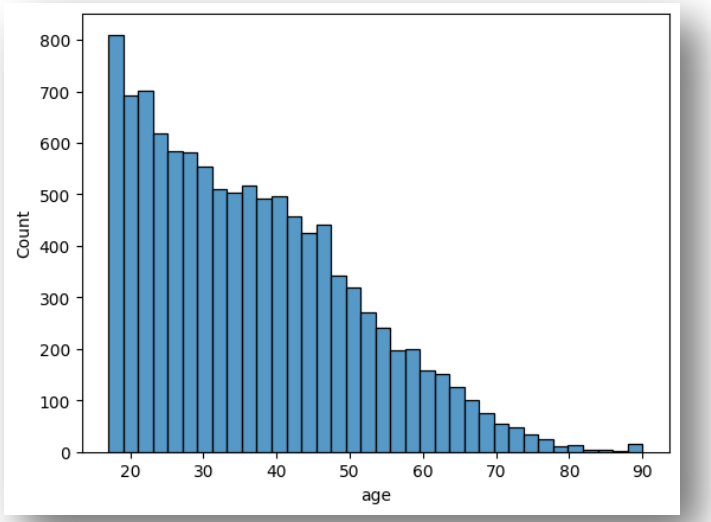
Bivariate analysis Numerical/Categorical



Age distribution of Male



Age distribution of Female



Bivariate analysis Categorical/Categorical

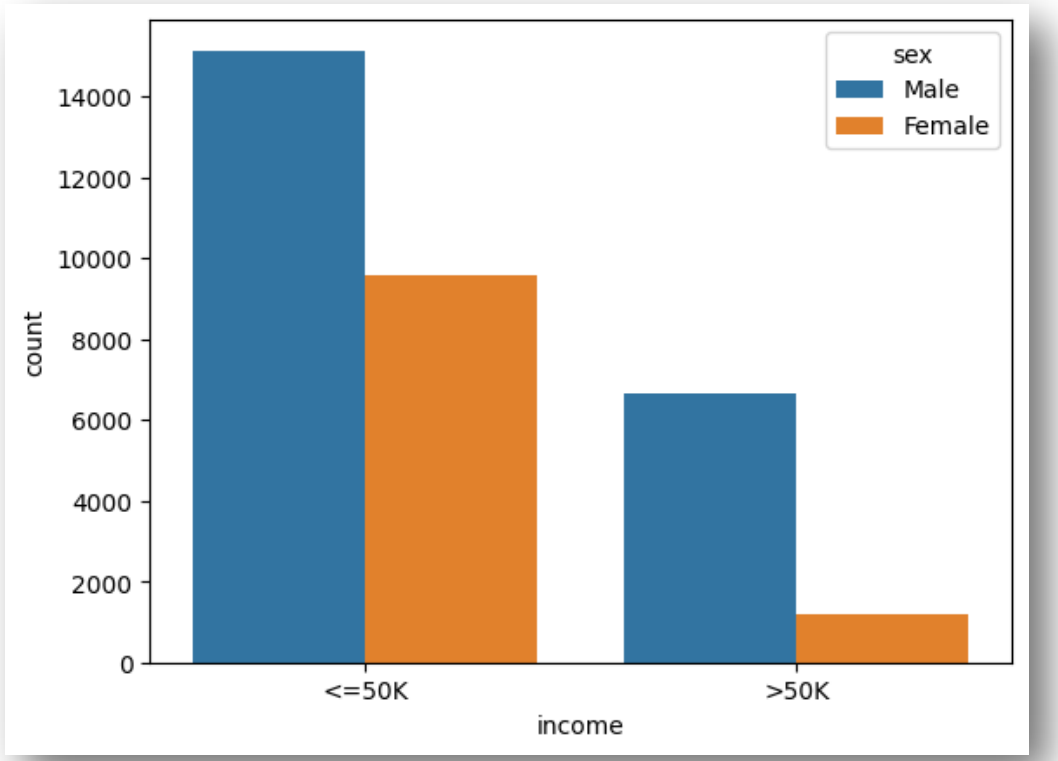
```
In [31]: # Plotting cross tabulation values for native-country and sex
         pd.crosstab(data['native-country'],data['sex'], margins=True)
```

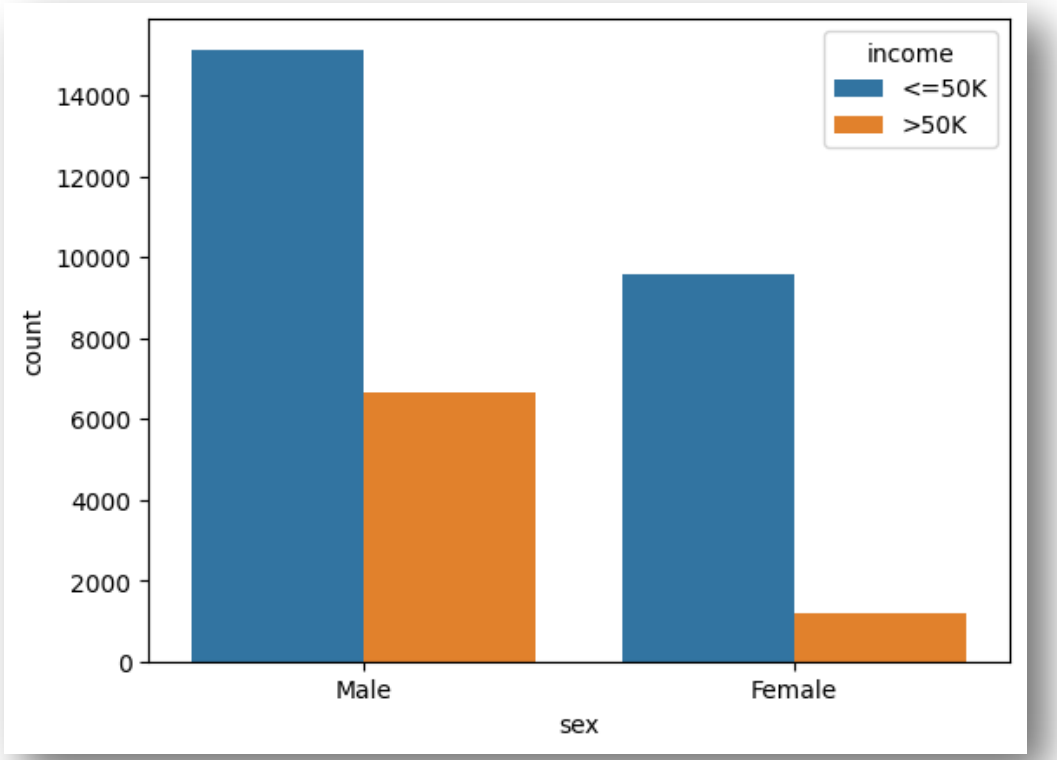
Out[31]:

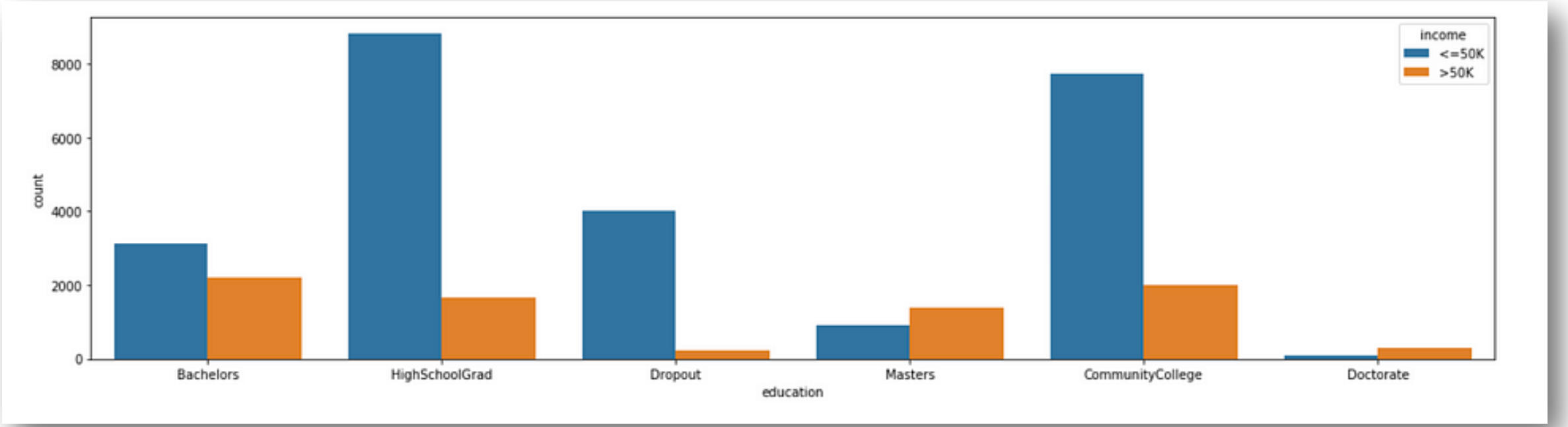
native-country	Female	Male	All
Others	1089	2302	3391
United-States	9682	19488	29170
All	10771	21790	32561

We look at two categorical attributes

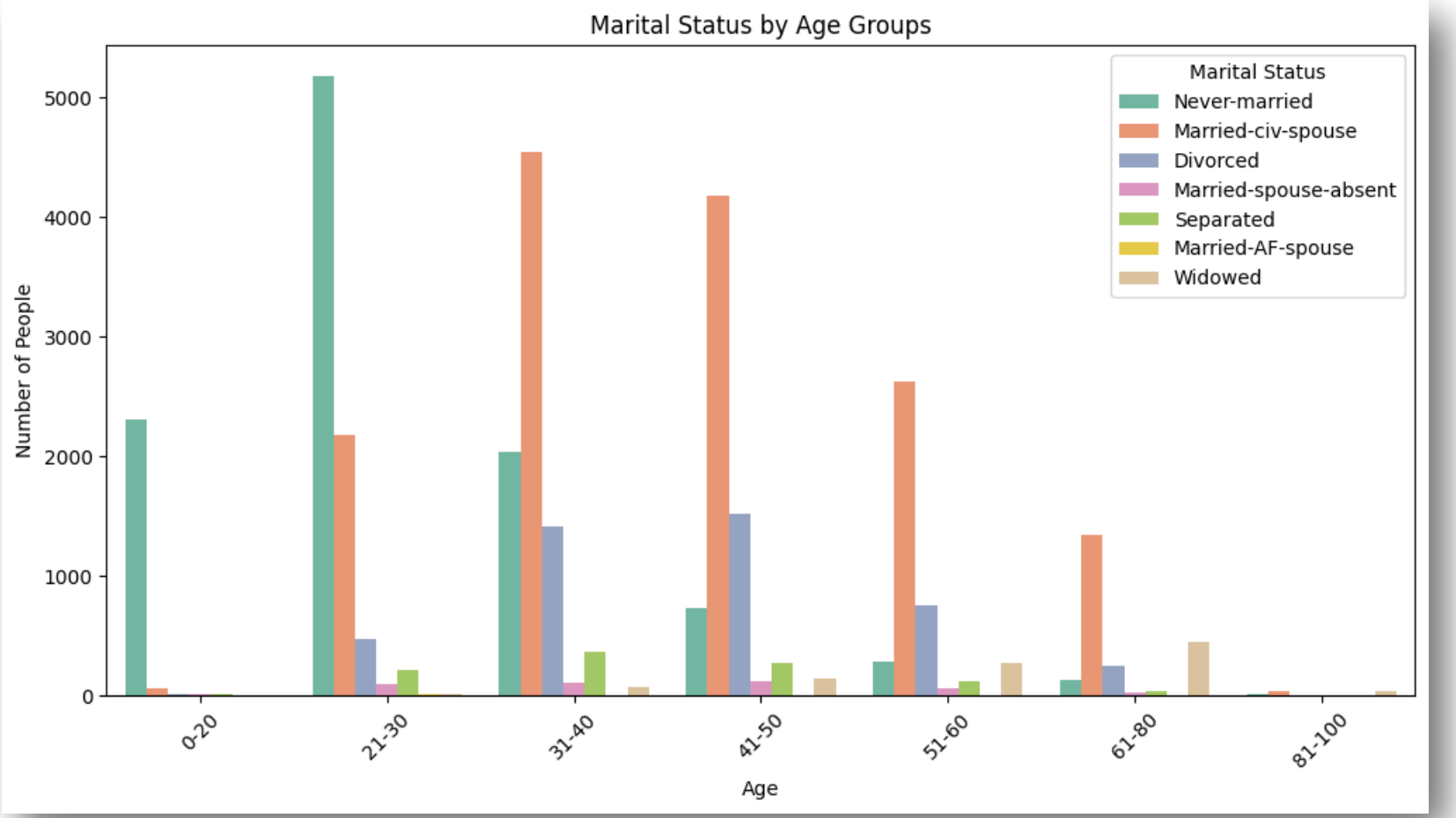
Same information, but would it matter what to put in a report?



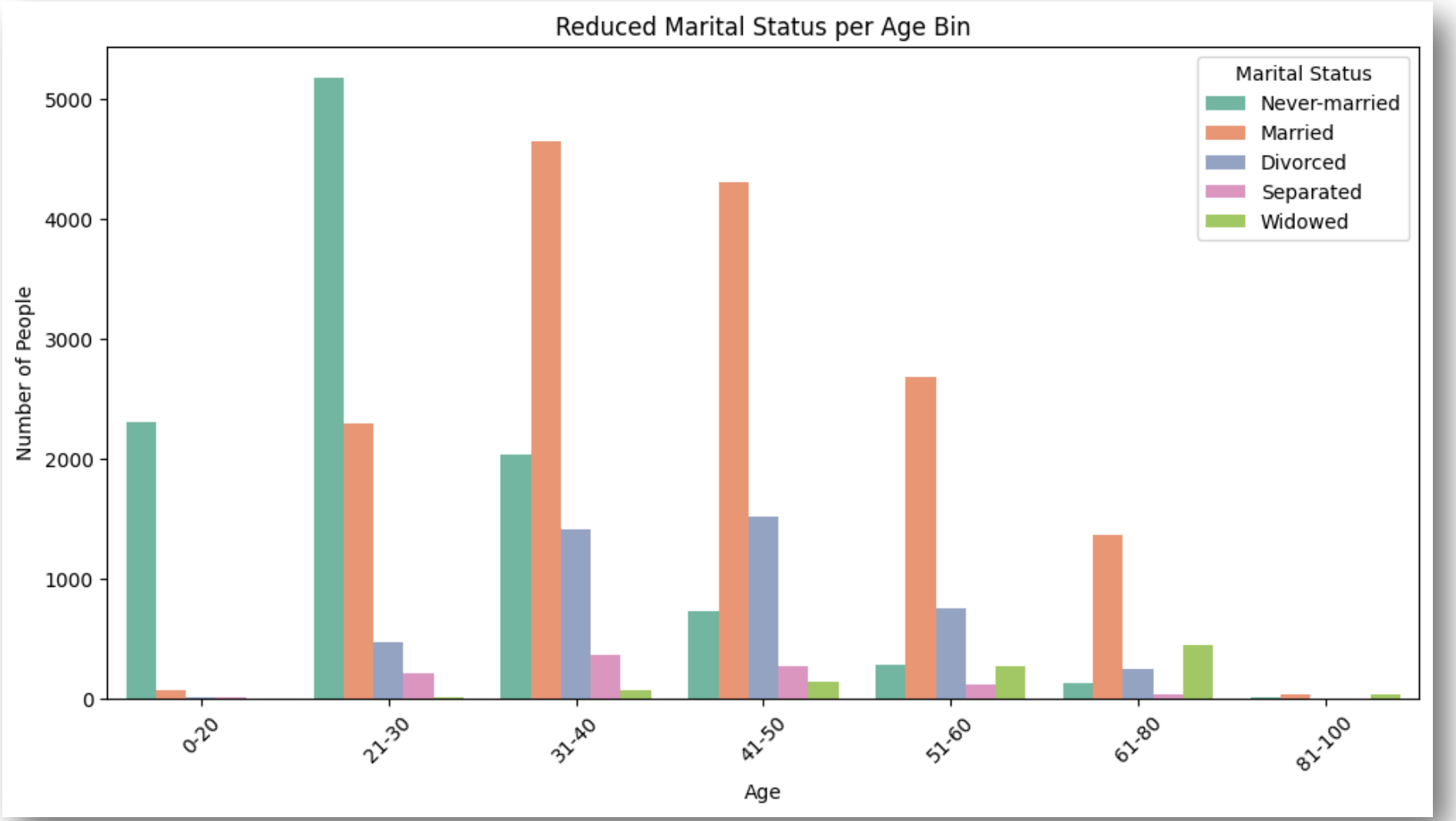



income
<=50K
>50K
count
8000
6000
4000
2000
0
Bachelors
HighSchoolGrad
Dropout
Masters
CommunityCollege
Doctorate
education

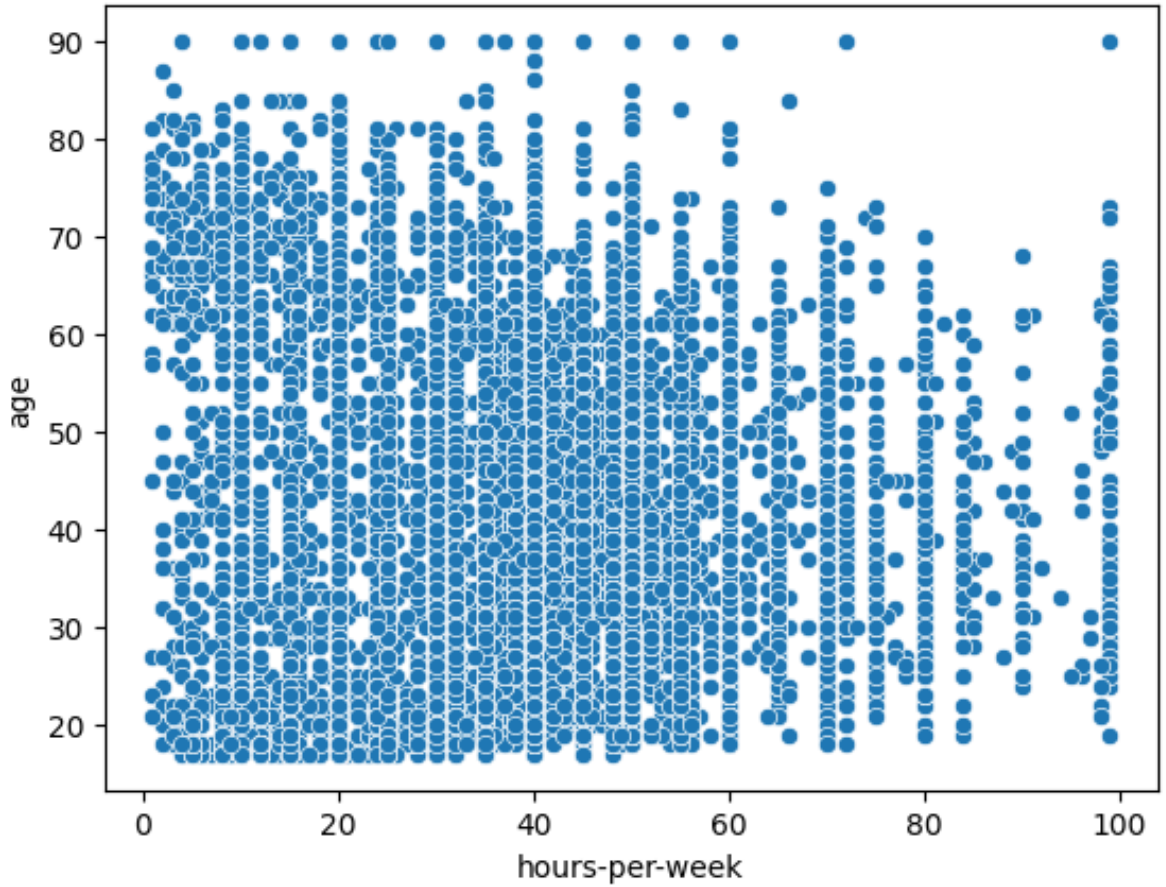
Bivariate analysis Categorical/Numerical

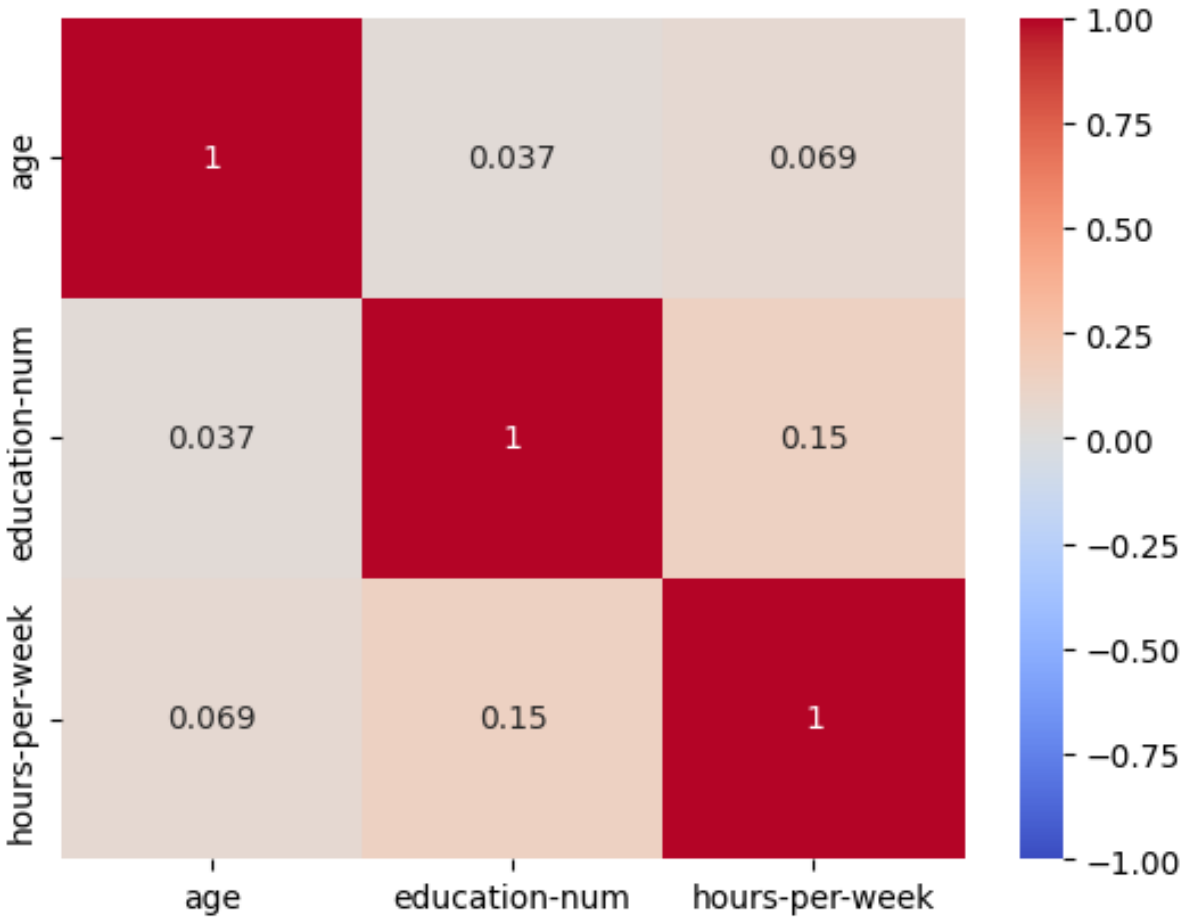


Category	Meaning / Explanation
Never-married	The person has **never legally married**.
Married-civ-spouse	Married **with a civil/legal spouse** (living together).
Married-AF-spouse	Married **with an Armed Forces spouse** (military).
Married-spouse-absent	Legally married but the **spouse is not present** (e.g., separated, working elsewhere).
Separated	Legally married but **currently separated** from the spouse.
Divorced	Legally **divorced** from a spouse.
Widowed	Spouse has **died**.


Reduced Marital Status per Age Bin
Marital Status
Never-married
Married
Divorced
Separated
Widowed
Number of People
0
1000
2000
3000
4000
5000
0-20
21-30
31-40
41-50
51-60
61-80
81-100
Age

Bivariate analysis Numerical/Numerical





Nothing seems correlated

Assignment 1
Deliverables

1. Choose 2 datasets

The Kaggle site is a very interesting site to explore as it contains datasets for many tasks in data science and AI. You must select, from Kaggle, 2 datasets to explore.

For **Dataset 1**, it MUST be one among the 3 suggested below. The 3 suggestions are in 3 different domains: finance, healthcare, and mobile usage.

1. German Credit Dataset
   - Size: 10 columns, 1000 rows
   - Description: Financial data for credit scoring
   - Link

2. Heart Failure Prediction Dataset
   - Size: 12 columns, 918 rows
   - Description: Health-related data for heart failure prediction
   - Link

3. Mobile Device Usage Dataset
   - Size: 11 columns, 700 rows
   - Description: Technology/mobile behavior data
   - Link

For **Dataset 2**, you must select a dataset that is NOT within the 3 suggested sets above. Furthermore, your second dataset MUST be in a domain different from the first dataset. I want you to explore 2 different domains. Another constraint is to find a dataset with a minimum of 10 columns so you have various features to explore.

2. Report the story of each dataset

The purpose of the report (written within a Jupyter Notebook) is to illustrate 10 insights that you found through your analysis for each dataset. In the notebook, you will be able to alternate between text and code, both required.

Your Jupyter Notebook should include:

1. Group number, names and student numbers of group members
2. Introduction to provide the goal of the analysis/report and mention who the audience would be (imagine an audience who would read your report).
3. A description of the two datasets used *(see Dataset description requirement section)*
4. **A set of 10 insights for each dataset**. For EACH ONE:
    a. State the insight in a single sentence.
    b. Show supporting evidence from the data making sure that evidence is as self-explanatory as possible (graph with title, axis descriptions, etc)
    c. Mention what type of analysis was done to arrive at such insight *(Analysis description requirement section)*.
    d. Show how this evidence was obtained (and how to reproduce it – provide code)
5. Conclusion
6. References

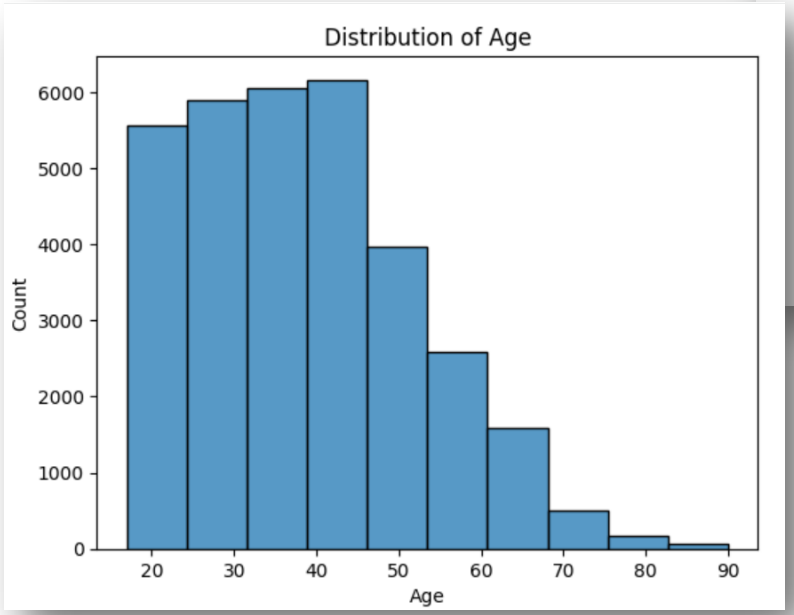
4. Analysis description requirements

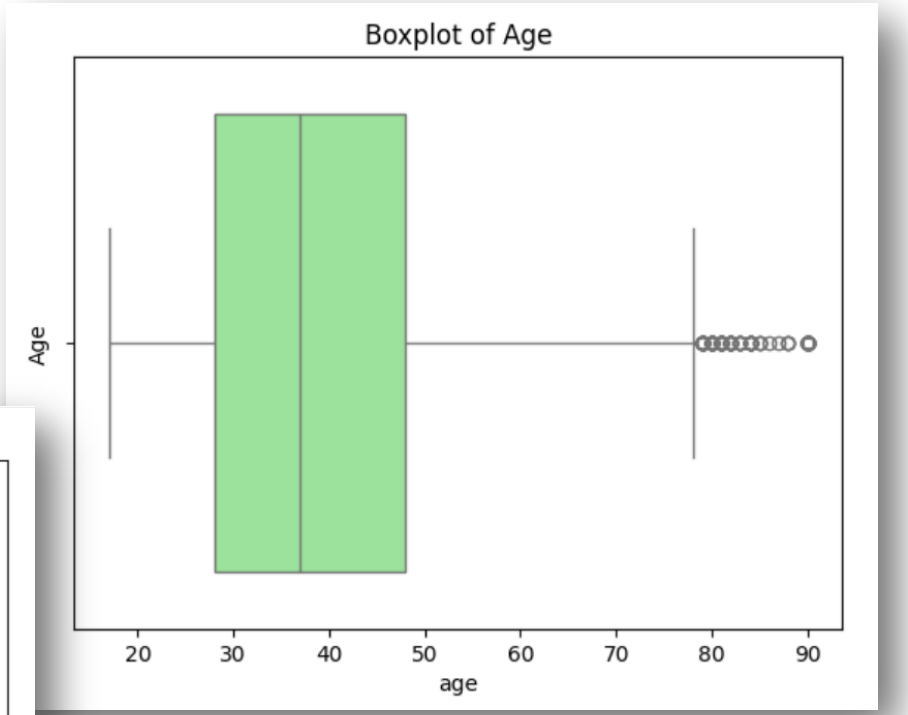
Given that we are in an **academic context**, I have additional requirements to make sure that you explore various types of analysis and visualization tools. Therefore, **among your 20 insights (10 for each dataset), you MUST have a diversity of supporting evidence including at least one of each of the analysis (written as r1 to r8) below for each dataset**:

a) Univariate analysis
   a. Numerical data:
      i. (r1) Simple histogram or boxplot for visualization of dispersion of a numerical variable
   b. Categorical data
      i. (r2) Countplot for a category with multiple values to see the distribution among values.
      ii. (r3) Grouped-Data countplot in which you group some values (and explain how you did the grouping).

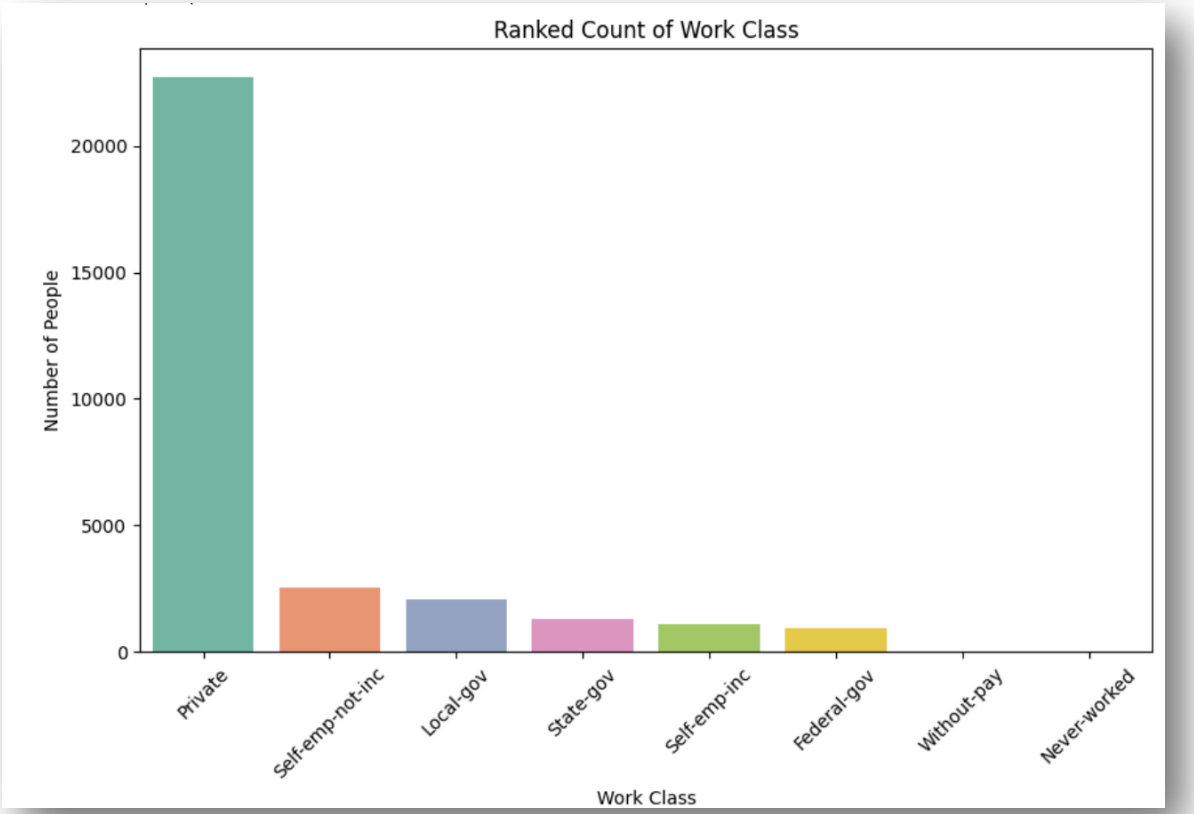
b) Bivariate analysis
   a. Numerical/Categorical
      i. (r4) Explore a numerical variable's distribution according to specific values of a categorical variable.
   b. Categorical/Categorical
      i. (r5) Comparing categories with 2 values
      ii. (r6) Comparing categories with more than 2 values for which you set the order (e.g. increasing counts, or alphabetical order)
   c. Categorical/Numerical
      i. (r7) Looking at a categorical variable's distribution among bins of a numerical feature. You should create the bins with specific values.
   d. Numerical/Numerical
      i. (r8) Use the scatterplot to highlight correlation.

R1 – Univariate Numerical



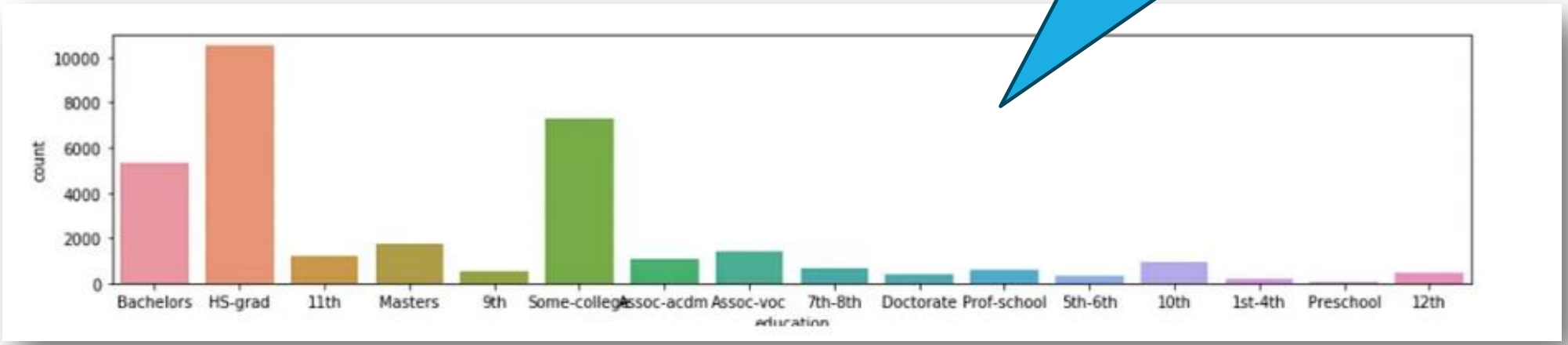


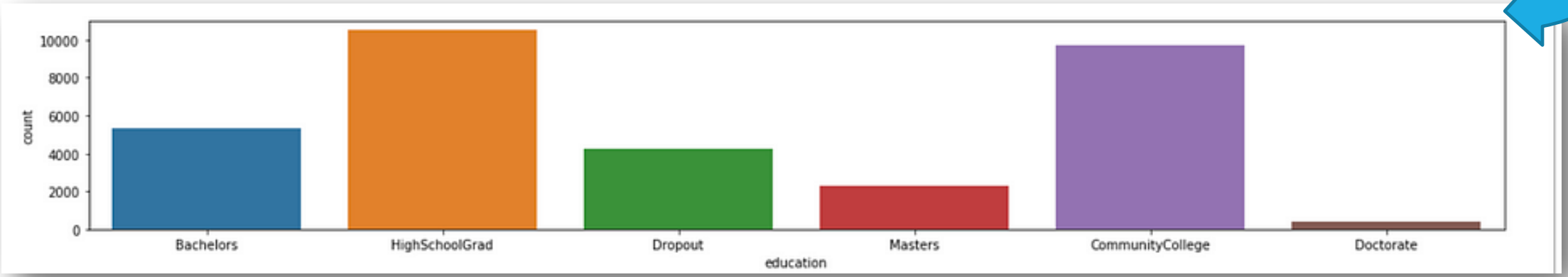
R2 – Univariate Categorical



R3 – Univariate Categorical - Grouped

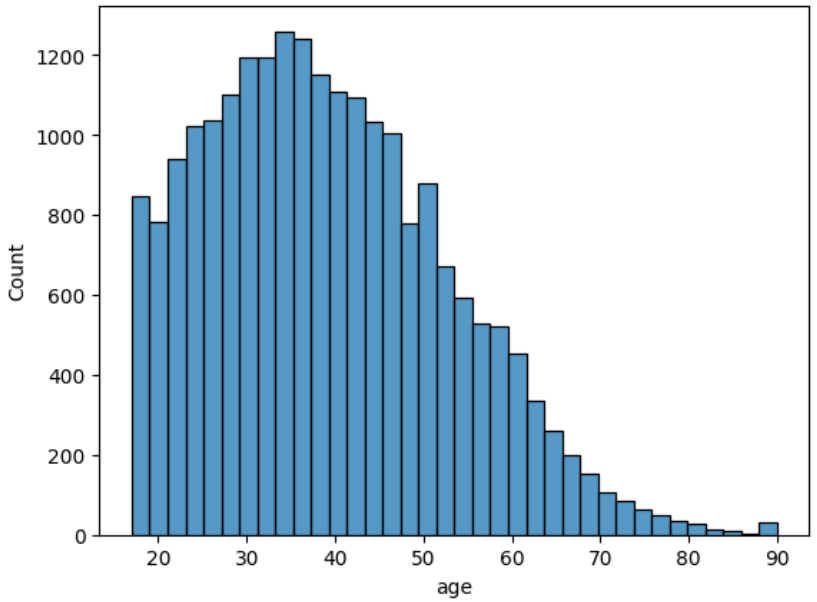
Provide domain justification for the grouping



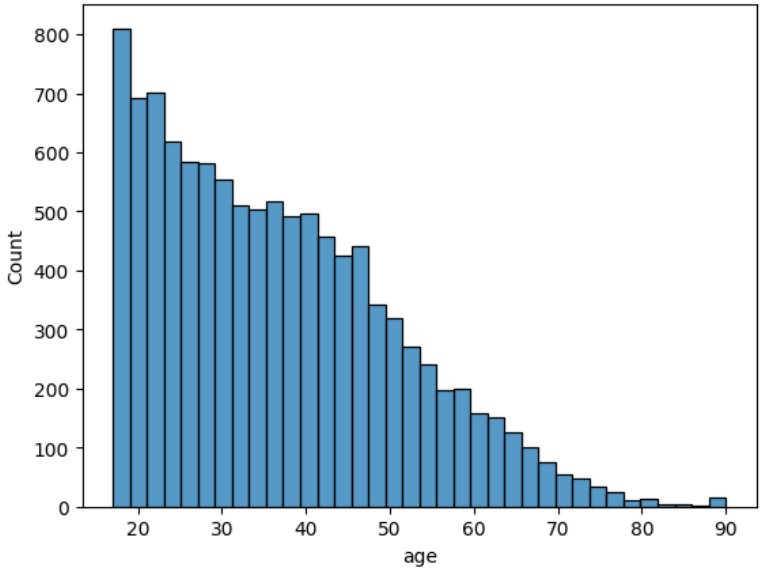


R4 – Bivariate Numerical/Categorical

Age distribution of Male

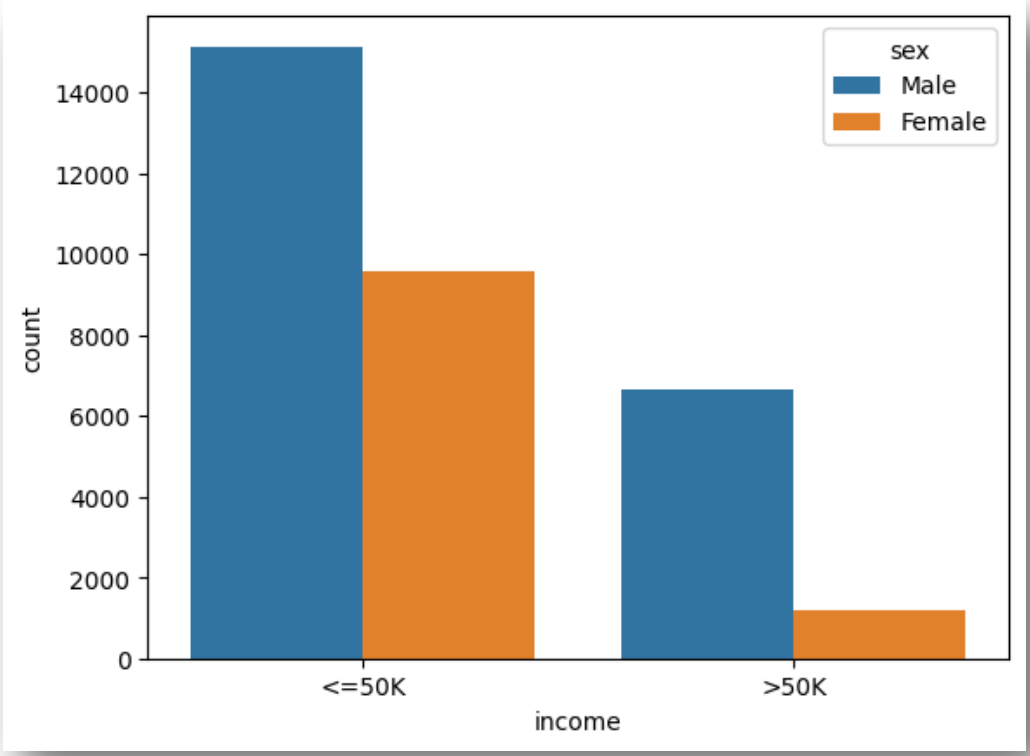


Age distribution of Female

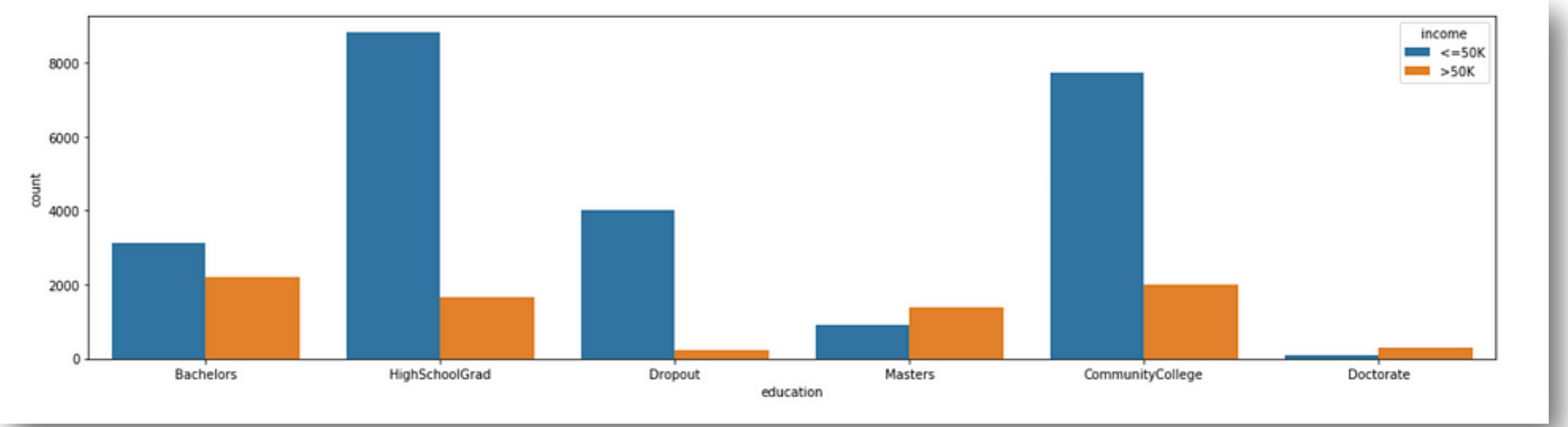


R5 – Bivariate Categorical/Categorical 2 values per category

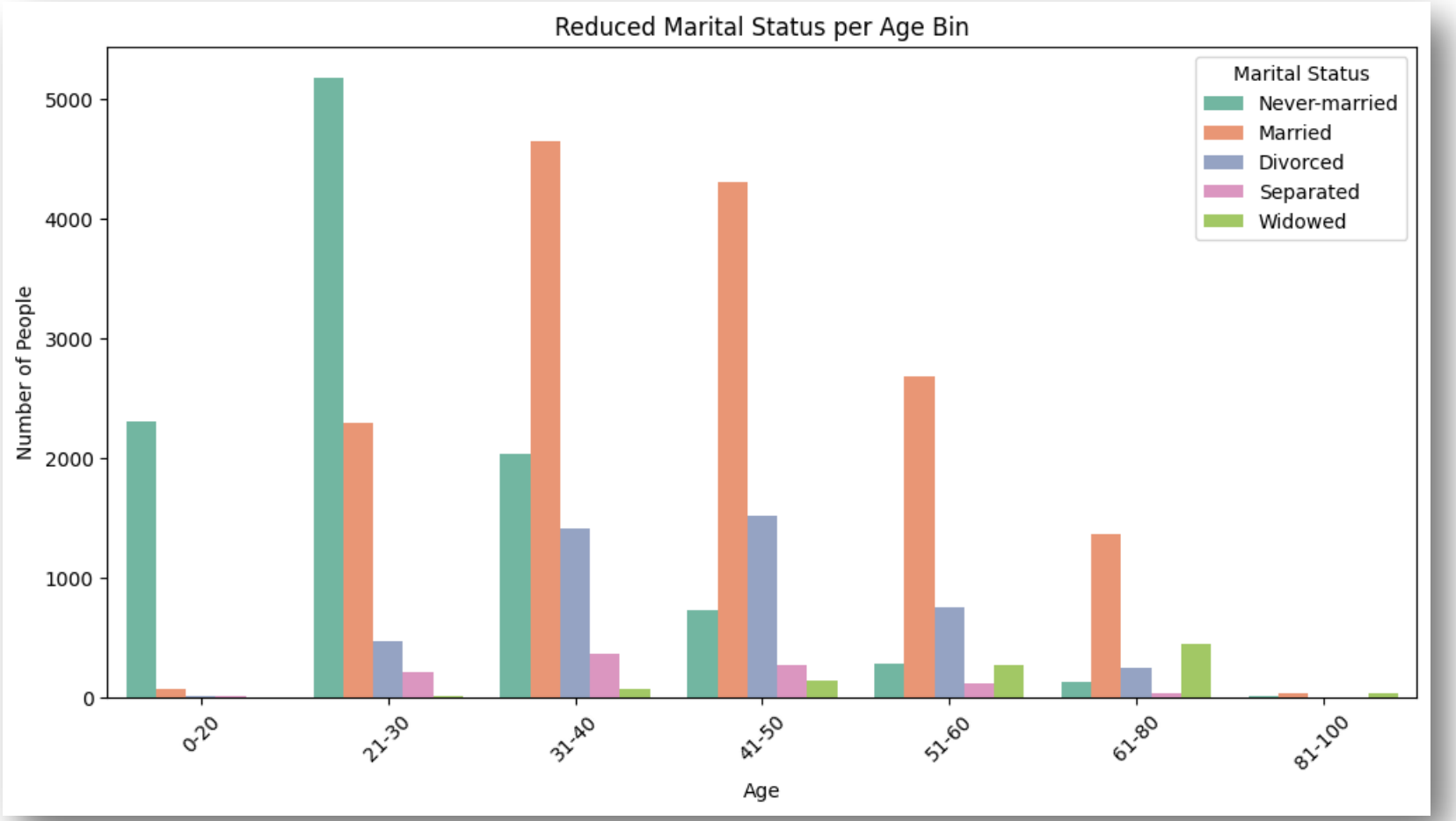
Explain why this view (and not flipped)



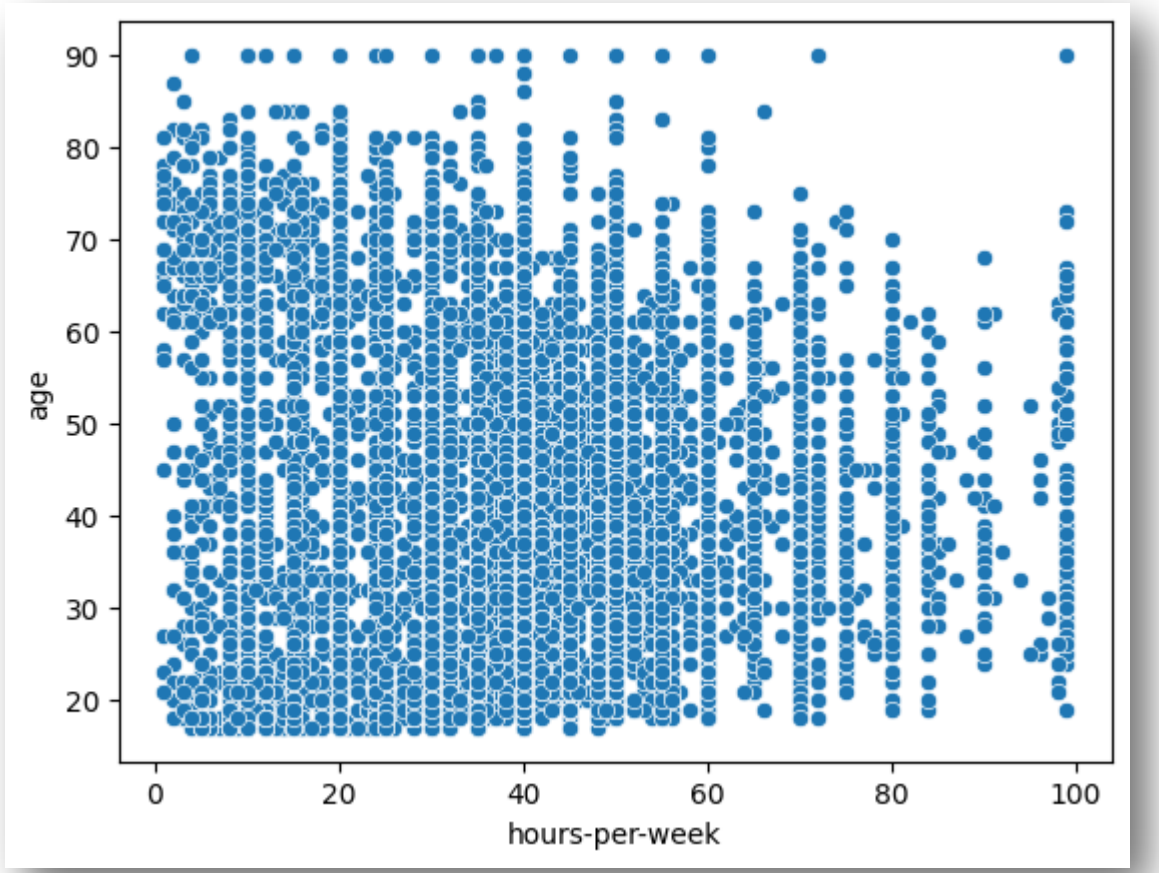
R6 – Bivariate Categorical/Categorical
More than 2 values for at least one category



R7 – Bivariate Categorical/Numerical



R8 – Bivariate Numerical/Numerical



Example of insights

Example 1

https://www.kaggle.com/datasets/nelgiriyewithana/billionaires-statistics-dataset/data

# rank The ranking of the billionaire in terms of wealth.	# finalWorth The final net worth of the billionaire in U.S. dollars.	category The category or industry in which the billionaire's business operates.	personName The full name of the billionaire.	# age The age of the billionaire.
1 2540	1000 211k	Finance & Invest... 14% Manufacturing 12% Other (1944) 74%	2638 unique values	18 101
1	211000	Fashion & Retail	Bernard Arnault & family	74
2	180000	Automotive	Elon Musk	51
3	114000	Technology	Jeff Bezos	59
4	107000	Technology	Larry Ellison	78
5	106000	Finance & Investments	Warren Buffett	92
6	104000	Technology	Bill Gates	67
7	94500	Media & Entertainment	Michael Bloomberg	81
8	93000	Telecom	Carlos Slim Helu & family	83
9	83400	Diversified	Mukesh Ambani	65
10	80700	Technology	Steve Ballmer	67
11	80500	Fashion & Retail	Francoise Bettencourt Meyers & family	69
12	79200	Technology	Larry Page	50

Key Features

- **rank**: The ranking of the billionaire in terms of wealth.
- **finalWorth**: The final net worth of the billionaire in U.S. dollars.
- **category**: The category or industry in which the billionaire's business operates.
- **personName**: The full name of the billionaire.
- **age**: The age of the billionaire.
- **country**: The country in which the billionaire resides.
- **city**: The city in which the billionaire resides.
- **source**: The source of the billionaire's wealth.
- **industries**: The industries associated with the billionaire's business interests.
- **countryOfCitizenship**: The country of citizenship of the billionaire.
- **organization**: The name of the organization or company associated with the billionaire.
- **selfMade**: Indicates whether the billionaire is self-made (True/False).
- **status**: "D" represents self-made billionaires (Founders/Entrepreneurs) and "U" indicates inherited or unearned wealth.
- **gender**: The gender of the billionaire.
- **birthDate**: The birthdate of the billionaire.
- **lastName**: The last name of the billionaire.
- **firstName**: The first name of the billionaire.
- **title**: The title or honorific of the billionaire.
- **date**: The date of data collection.
- **state**: The state in which the billionaire resides.
- **residenceStateRegion**: The region or state of residence of the billionaire.
- **birthYear**: The birth year of the billionaire.
- **birthMonth**: The birth month of the billionaire.
- **birthDay**: The birth day of the billionaire.
- **cpi_country**: Consumer Price Index (CPI) for the billionaire's country.
- **cpi_change_country**: CPI change for the billionaire's country.

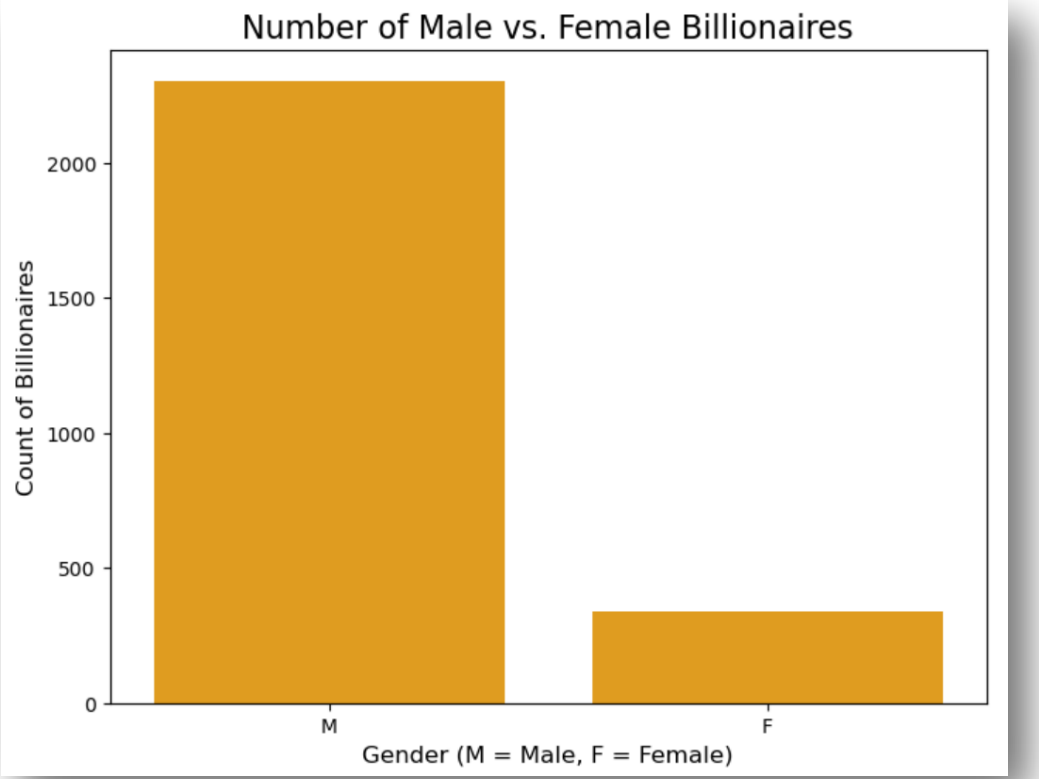
No need to use everything...

Insight 1

Insight: The large majority of billionaires are male, with women making up a small portion of the total number of billionaires.

Analysis Type: Countplot for a category with multiple values

How Evidence Was Obtained: Using the countplot function on the gender column to count the number of occurences of male and female billionaires in the dataset. With this we can visualize the disparity between the genders.

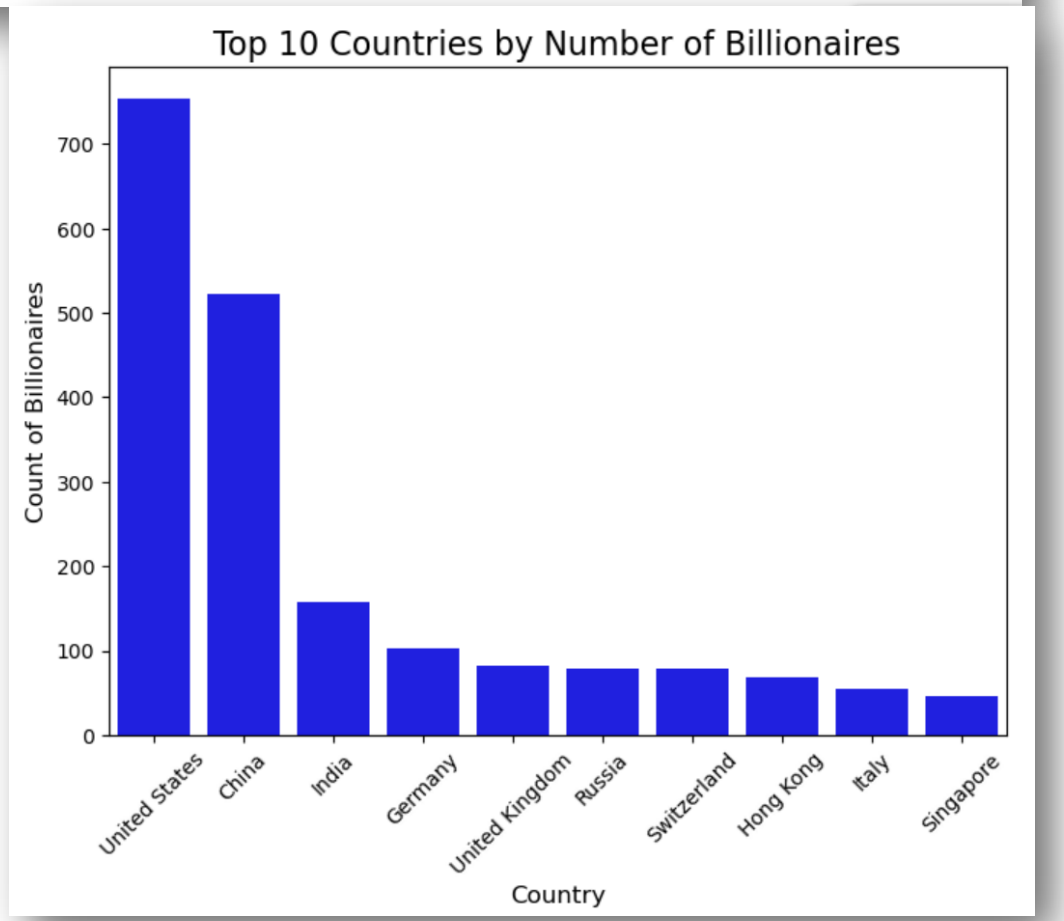


Insight 3

Insight: The U.S. and China dominate the billionaire list, with vastly more billionaires than any other country.

Analysis Type: Countplot for a category with multiple values

How Evidence Was Obtained: We first get the count of the number of billionaires in each country with the value_counts() function. Then, we get the top 10 using the head(10) function. We can then visualize in descending order the number of billionaires by country.



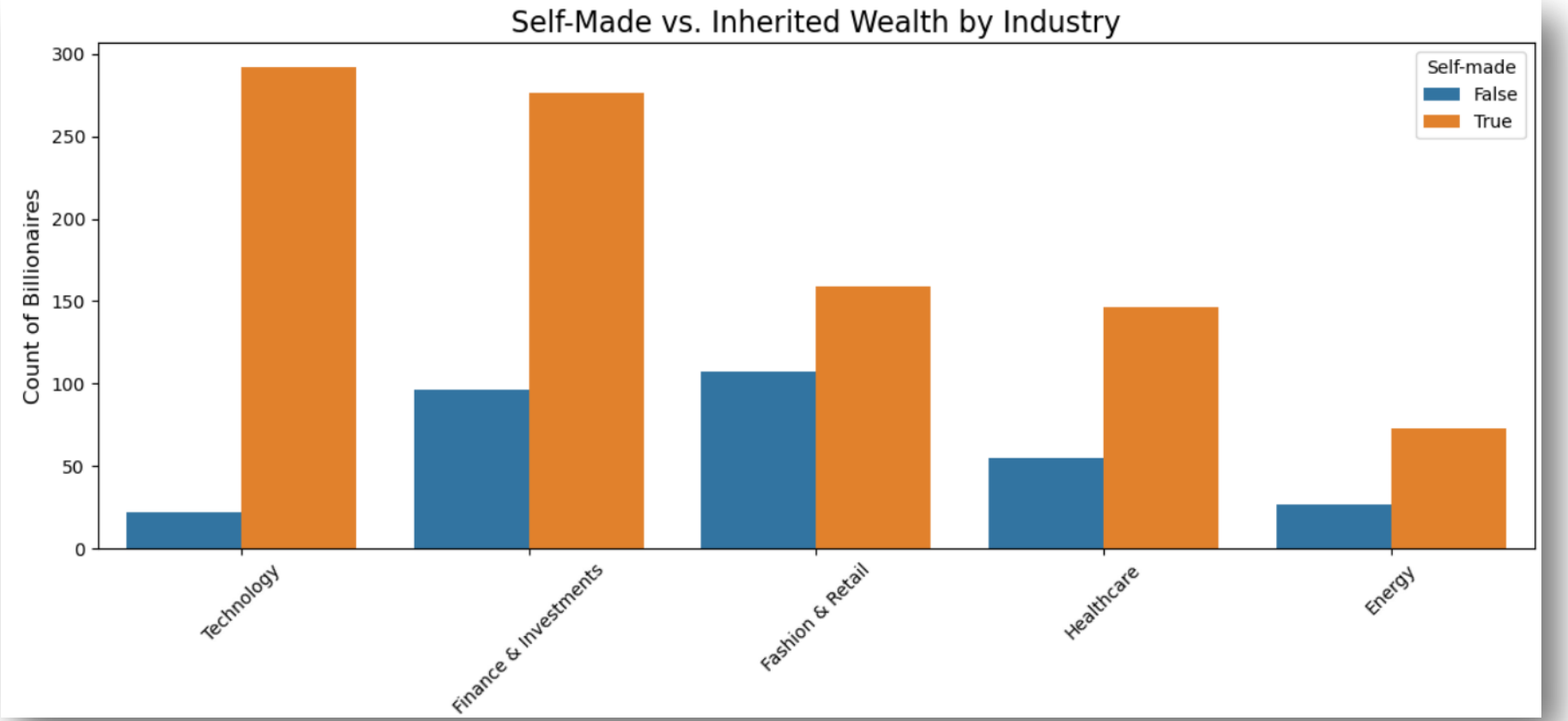
Insight 10

Insight: The technology industry creates the most self-made billionaires because it has a low barrier of entry and it is easy to grow big quickly.

Analysis Type: Compare category with 2 values against category with more than 2 values

How Evidence Was Obtained: This countplot compares the self-made versus inherited status for billionaires across different industries, focusing on technology and other major sectors. It shows the dominance of self-made billionaires in the tech industry.

That's all in the data?

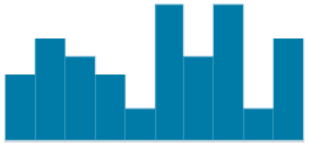
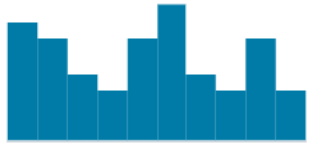
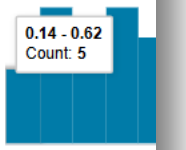


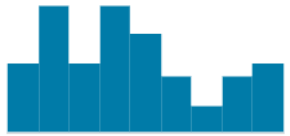
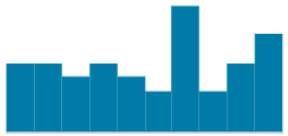
Example 2

Agriculture and Farming Dataset

Insights and Trends in Agriculture and Farming: A Comprehensive Data Analysis

https://www.kaggle.com/datasets/bhadramohit/agriculture-and-farming-dataset

Farm_ID FarmID	Crop_Type Type of Crop used	Farm_Area(acres) Area of Farm land	Irrigation_Type Type of Irrigation applied	Fertilizer_Used(to... Fertilizers used	Pesticide_ Pesticides us
50 unique values	Cotton 14% Barley 14% Other (36) 72%	 12.5 484	Drip 30% Flood 26% Other (22) 44%	 0.5 9.96	 0.14
F001	Cotton	329.4	Sprinkler	8.14	2.21
F002	Carrot	18.67	Manual	4.77	4.36
F003	Sugarcane	306.03	Flood	2.91	0.56
F004	Tomato	380.21	Rain-fed	3.32	4.35
F005	Tomato	135.56	Sprinkler	8.33	4.48
F006	Sugarcane	12.5	Sprinkler	6.42	2.25
F007	Soybean	360.06	Drip	1.83	2.37
F008	Rice	464.6	Drip	5.18	0.91

# Pesticide_Used(kg) Pesticides used	# Yield(tons) Yielding	Soil_Type Type of soil on which it is yield	Season Season	# Water_Usage(cub... Usage of Water
 0.14 4.99	 3.86 48	Clay 28% Loamy 22% Other (25) 50%	Zaid 46% Kharif 32% Other (1) 22%	
2.21	14.44	Loamy	Khar	
4.36	42.91	Peaty	Khar	
0.56	33.44	Silty	Khar	
4.35	34.08	Silty	Zaid	
4.48	43.28	Clay	Zaid	
2.25	38.18	Loamy	Zaid	
2.37	44.93	Sandy	Rabi	
0.91	4.23	Silty	Khar	
4.93	3.86	Peaty	Rabi	
2.67	17.25	Sandy	Khar	
1.24	32.85	Clay	Zaid	

what do zaid, kharif and rabi refer to as seasons

In the context of **agriculture**, **Zaid, Kharif, and Rabi** refer to **three distinct crop-growing seasons** in South Asia:

Season	Time of Year	Notes / Crops
Rabi	Winter	Sown after the monsoon, around **October–December**, harvested in **March–April**. Examples: wheat, barley, mustard.
Kharif	Monsoon / Rainy	Sown at the **start of the rainy season**, around **June–July**, harvested in **September–October**. Examples: rice, maize, cotton.
Zaid	Summer / Short-season	Crops grown in the **short season between Rabi and Kharif**, around **March–June**. Examples: watermelon, cucumber, summer vegetables.

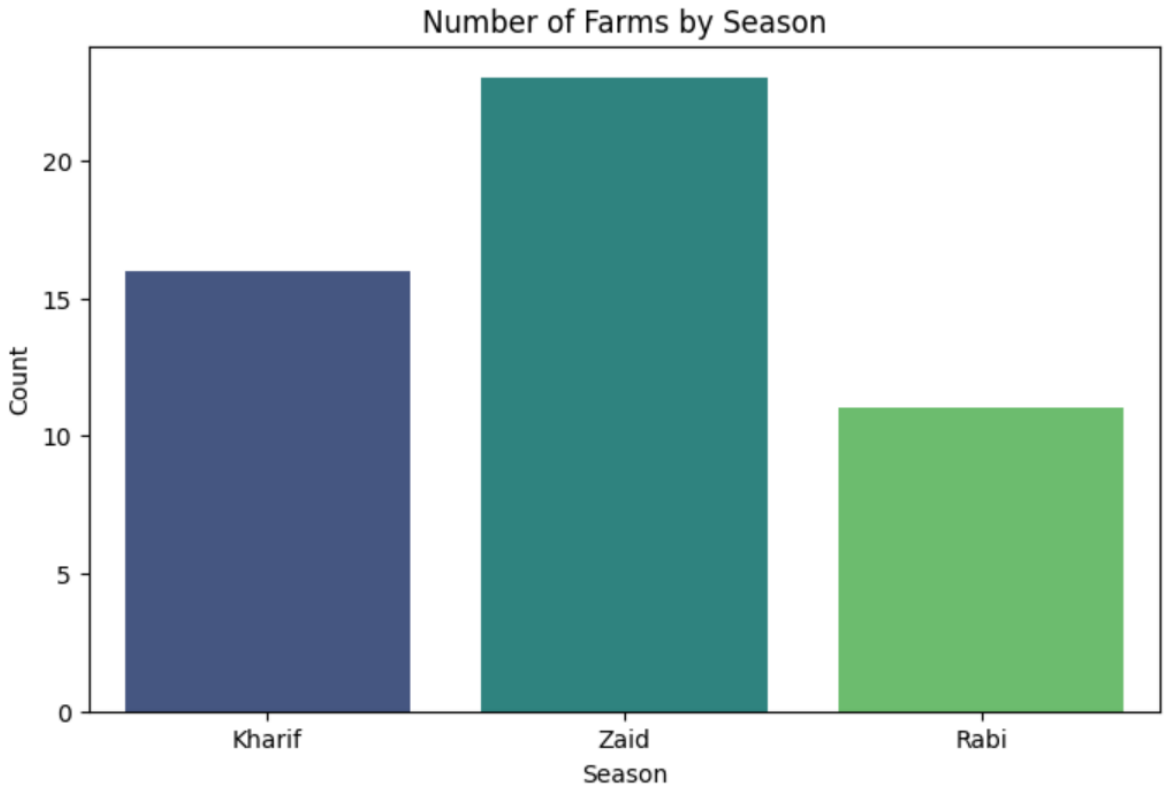
Insight 2: Zaid is the most common farming season.

Supporting Evidence: A countplot of different seasons shows that Kharif dominates.

Analysis Type: Univariate Analysis (Categorical) → Countplot.

Be careful with your explanation… there is a contradiction here.

Be more specific in your title (active farms?)



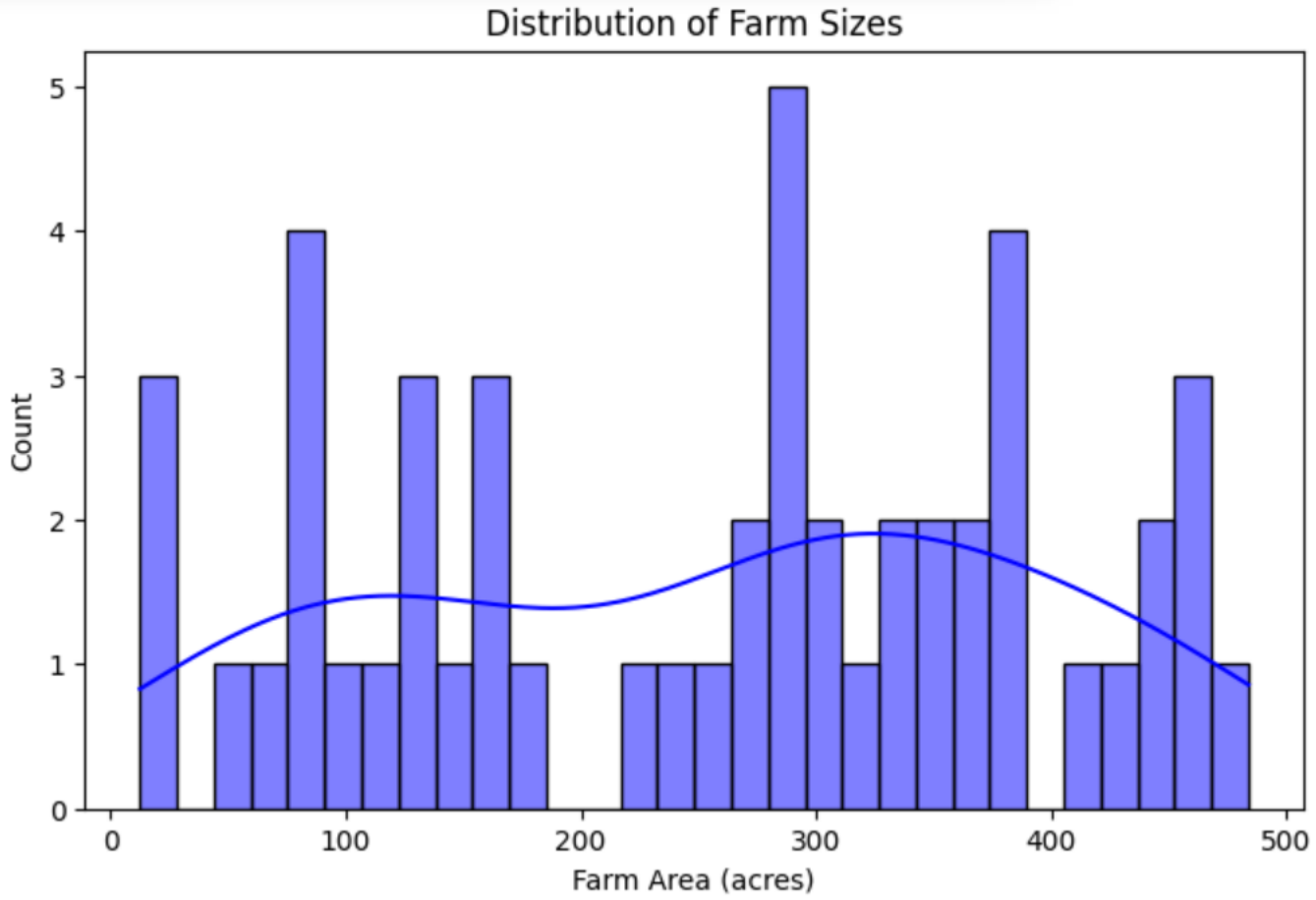
Insight 1: The distribution of farm sizes is right-skewed, with many small farms and fewer large farms.

Supporting Evidence: A histogram of farm sizes shows most farms are smaller, with a few much larger ones.

Analysis Type: Univariate Analysis (Numerical) → Histogram

It's actually left-skewed (long tail on the left)

Not that obvious... Perhaps say something else?

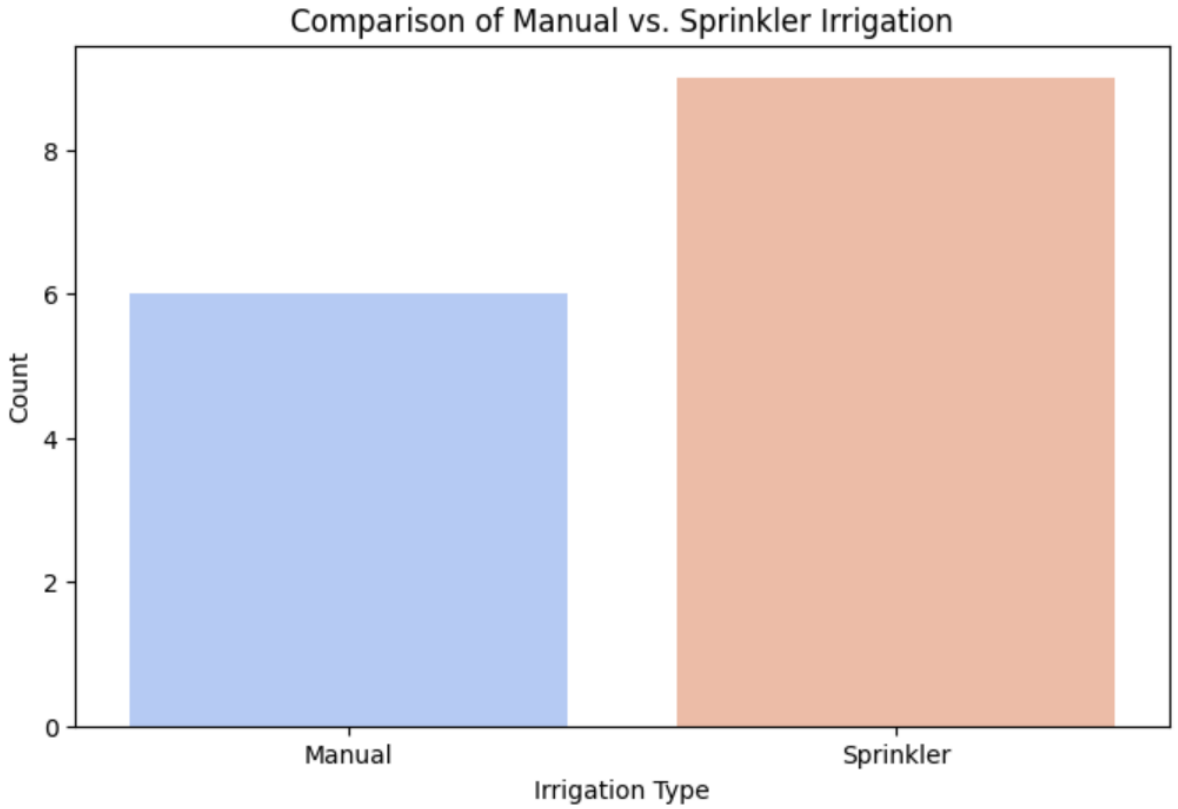


Insight 4: Sprinkler irrigation is more common than manual irrigation.

Supporting Evidence: A comparison of two irrigation types.

Analysis Type: Bivariate Analysis (Categorical/Categorical) → Compare Two Categories

This is a UNIVARIATE analysis… two values of the same variable

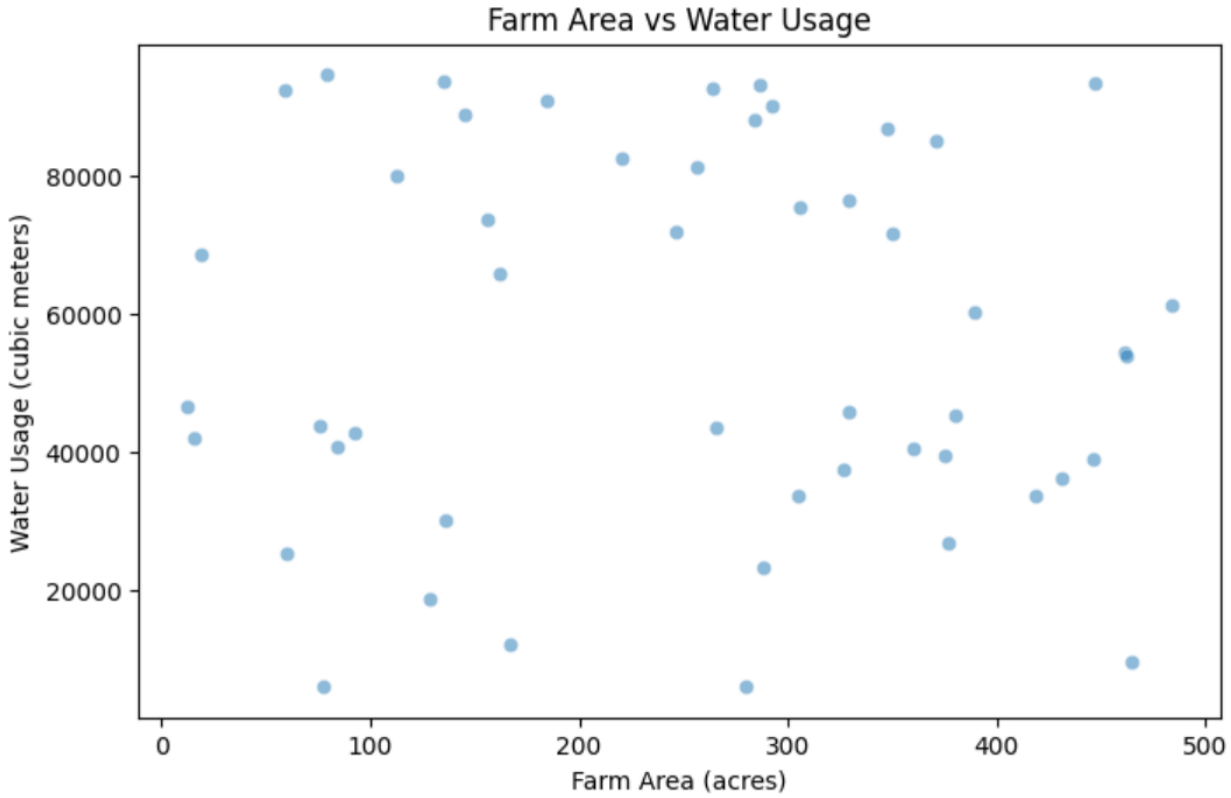


Insight 6: There is a moderate correlation between farm size and water usage.

Supporting Evidence: A scatter plot of farm area vs. water usage shows an increasing trend.

Analysis Type: Bivariate Analysis (Numerical/Numerical) → Correlation Scatterplot

Really?

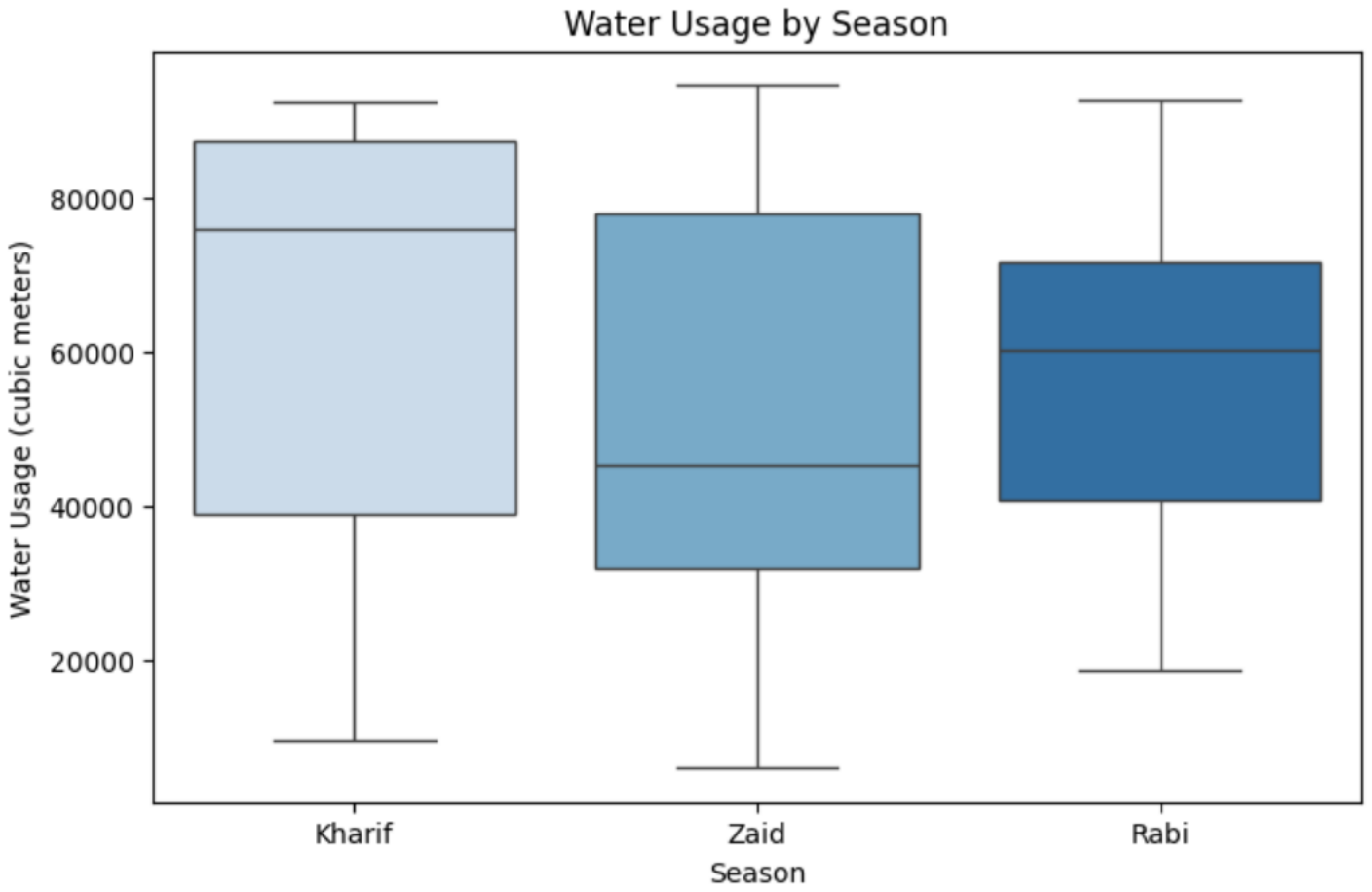


Insight 9: Farms in Kharif season use significantly more water than those in other seasons.

Supporting Evidence: A box plot of water usage grouped by season.

Analysis Type: Bivariate Analysis (Numerical/Categorical) → Boxplots per value

OK...

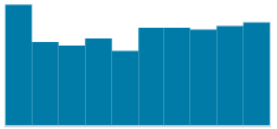
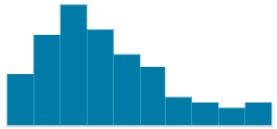
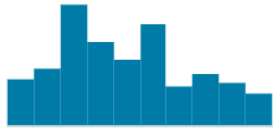
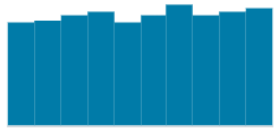


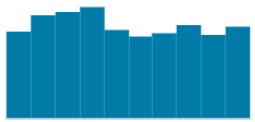
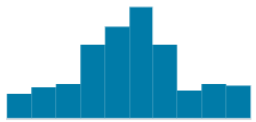
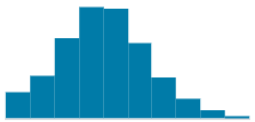
Example 3

Gym Members Exercise Dataset

Analyzing Fitness Patterns and Performance Across Diverse Gym Experience Levels

https://www.kaggle.com/datasets/valakhorasani/gym-members-exercise-dataset

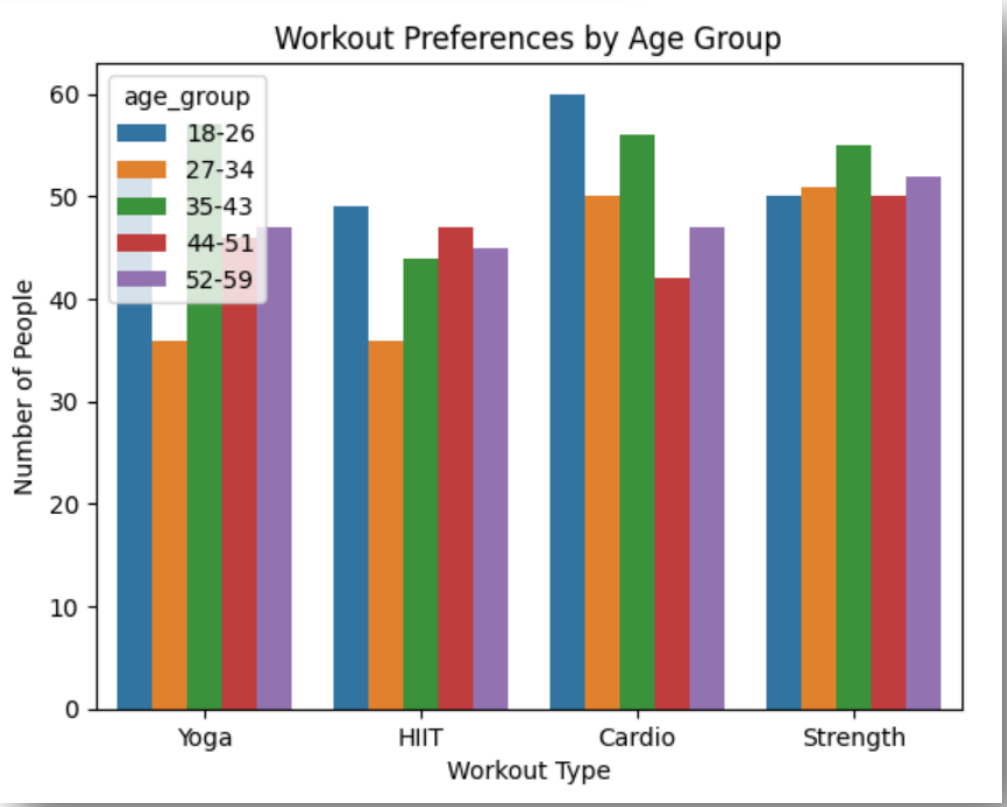
# Age The age of the gym member (in years)	Gender The gender of the gym member (Male or Female)	# Weight (kg) Member's body weight measured in kilograms	# Height (m) Member's height measured in meters	# Max_BPM Maximum beats per minute (BPM) recorded during a workout session
 18 59	Male 53% Female 47%	 40 130	 1.5 2	 160 199
56	Male	88.3	1.71	180
46	Female	74.9	1.53	179
32	Female	68.1	1.66	167
25	Male	53.2	1.7	190
38	Male	46.1	1.79	188
56	Female	58.0	1.68	168
36	Male	70.3	1.72	174
40	Female	69.7	1.51	189
28	Male	121.7	1.94	185
28	Male	101.8	1.84	169
41	Male	120.8	1.67	188
53	Male	51.7	1.7	175

# Avg_BPM Average beats per minute (BPM) during the workout session	# Resting_BPM Member's heart rate (BPM) before starting the workout, at rest	# Session_Duration... The total time spent during a workout session, measured in hours	# Calories_Burned The number of calories burned during the workout session	Workout_Type The type of workout performed (e.g., Cardio, Strength, Yoga, HIIT)
 120 169	 50 74	 0.5 2	 303 1.78k	Strength 27% Cardio 26% Other (460) 47%
157	60	1.69	1313.0	Yoga
151	66	1.3	883.0	HIIT
122	54	1.11	677.0	Cardio
164	56	0.59	532.0	Strength
158	68	0.64	556.0	Strength
156	74	1.59	1116.0	HIIT
169	73	1.49	1385.0	Cardio
141	64	1.27	895.0	Cardio
127	52	1.03	719.0	Strength
136	64	1.08	808.0	Cardio
146	54	0.82	593.0	HIIT
152	72	1.15	865.0	HIIT

Insight:

All age groups like doing Strength exercises equally, while Yoga and HIIT remains less popular, especially for people in age 27-34; and Cardio is more popular among younger people (18-43)

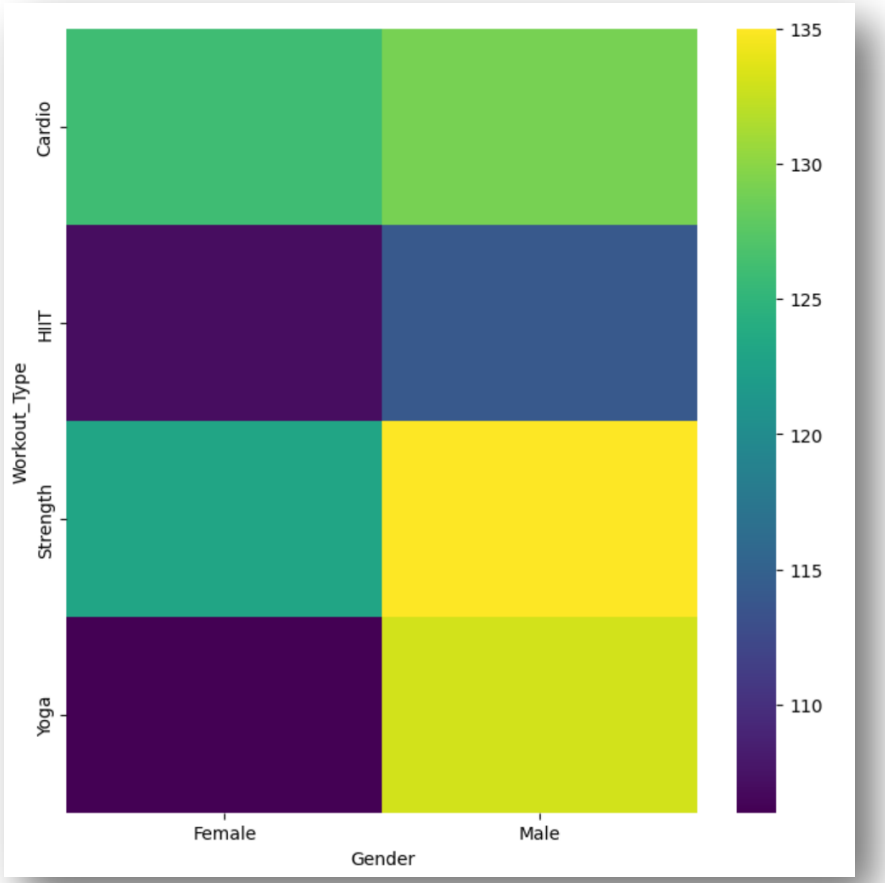
Please be careful with hiding the graph



Insight:

Yoga and Strength is most popular for Men, while Woman prefer doing Cardio

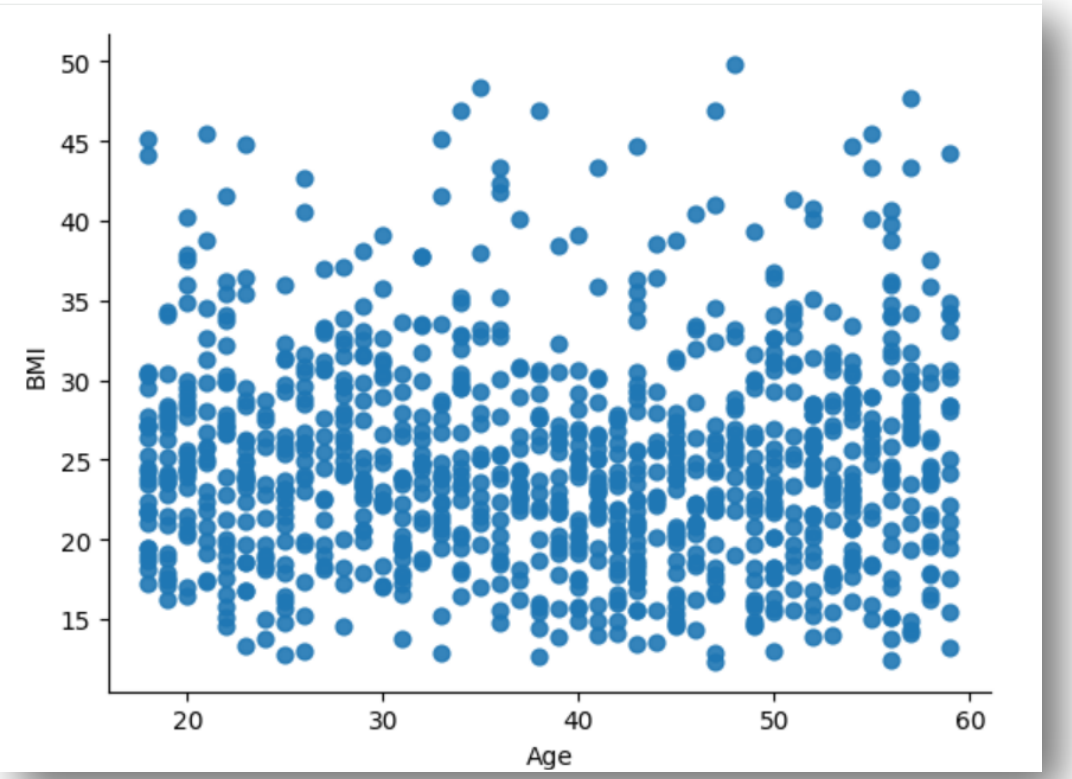
OK if you want to do something different, but still, the insight should be interpreted easily from the evidence provided.



Insight:

BMI level, an indication for obesity and healthiness, remains quite evenly distributed among all age groups, with people have high BMI slightly te ds to be in 18-24 and 55-60.

If including additional info… make sure there are references for it



EXPLORATORY DATA ANALYSIS

- Case Study 2 – Census data
- Assignment 1 Examples