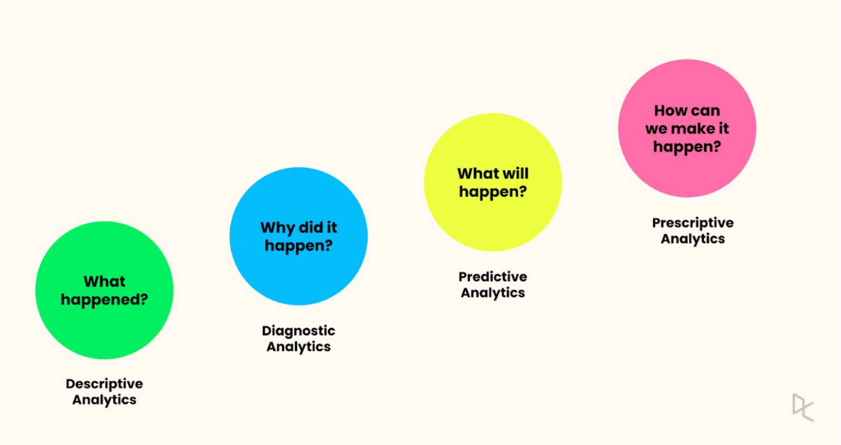
Shopping Dataset Case Study

THE KNOWLEDGE HOUSE

Agenda - Schedule

1. Case Study Introduction
2. Data Visualizations
3. Break (30 Mins)
4. Continue Case Study



Agenda - Goals

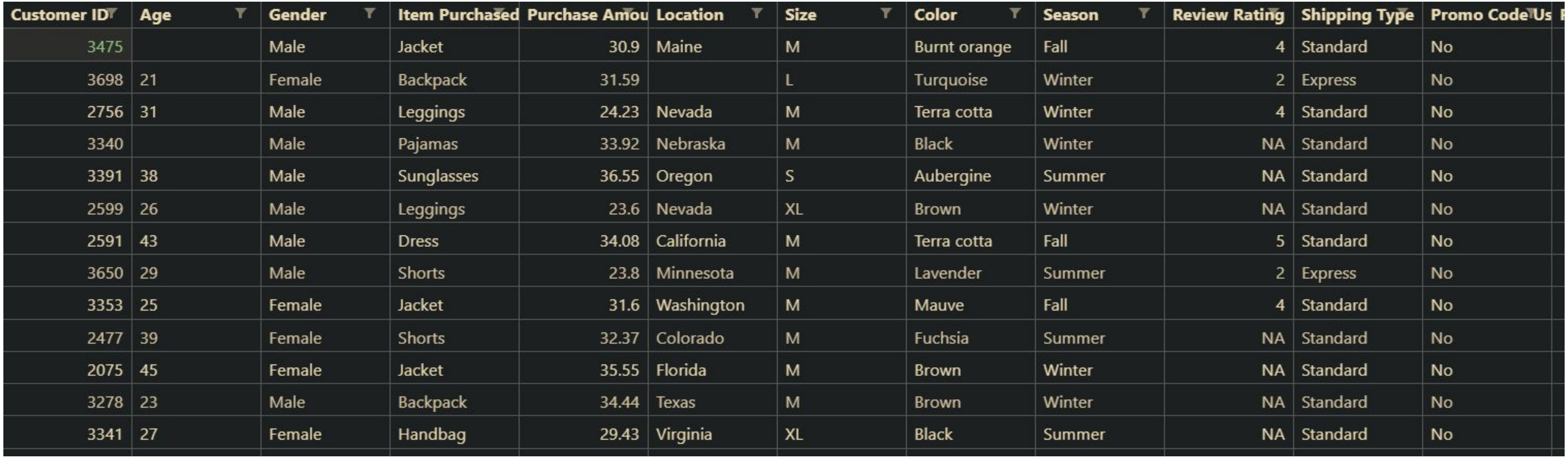
- Apply basic and intermediate pandas methods to **explore a structured dataset**
- Perform **univariate and bivariate analysis** on real-world shopping data
- Create visualizations using seaborn to support your findings
- Use grouping and aggregation techniques such as groupby(), pivot_table(), qcut(), and agg()
- Develop and **communicate insights clearly based on observed data patterns**, not just the code used

Announcements

- Week 8 Pre-Class Quiz due 4/29 (2 *attempts*)
- Review Session on 5/1
- TLAB #3 due 5/14

"*be-leaf in yourself*!"

Shopping Dataset Case Study



Customer ID	Age	Gender	Item Purchased	Purchase Amou	Location	Size	Color	Season	Review Rating	Shipping Type	Promo Code Us
3475		Male	Jacket	30.9	Maine	M	Burnt orange	Fall	4	Standard	No
3698	21	Female	Backpack	31.59		L	Turquoise	Winter	2	Express	No
2756	31	Male	Leggings	24.23	Nevada	M	Terra cotta	Winter	4	Standard	No
3340		Male	Pajamas	33.92	Nebraska	M	Black	Winter	NA	Standard	No
3391	38	Male	Sunglasses	36.55	Oregon	S	Aubergine	Summer	NA	Standard	No
2599	26	Male	Leggings	23.6	Nevada	XL	Brown	Winter	NA	Standard	No
2591	43	Male	Dress	34.08	California	M	Terra cotta	Fall	5	Standard	No
3650	29	Male	Shorts	23.8	Minnesota	M	Lavender	Summer	2	Express	No
3353	25	Female	Jacket	31.6	Washington	M	Mauve	Fall	4	Standard	No
2477	39	Female	Shorts	32.37	Colorado	M	Fuchsia	Summer	NA	Standard	No
2075	45	Female	Jacket	35.55	Florida	M	Brown	Winter	NA	Standard	No
3278	23	Male	Backpack	34.44	Texas	M	Brown	Winter	NA	Standard	No
3341	27	Female	Handbag	29.43	Virginia	XL	Black	Summer	NA	Standard	No

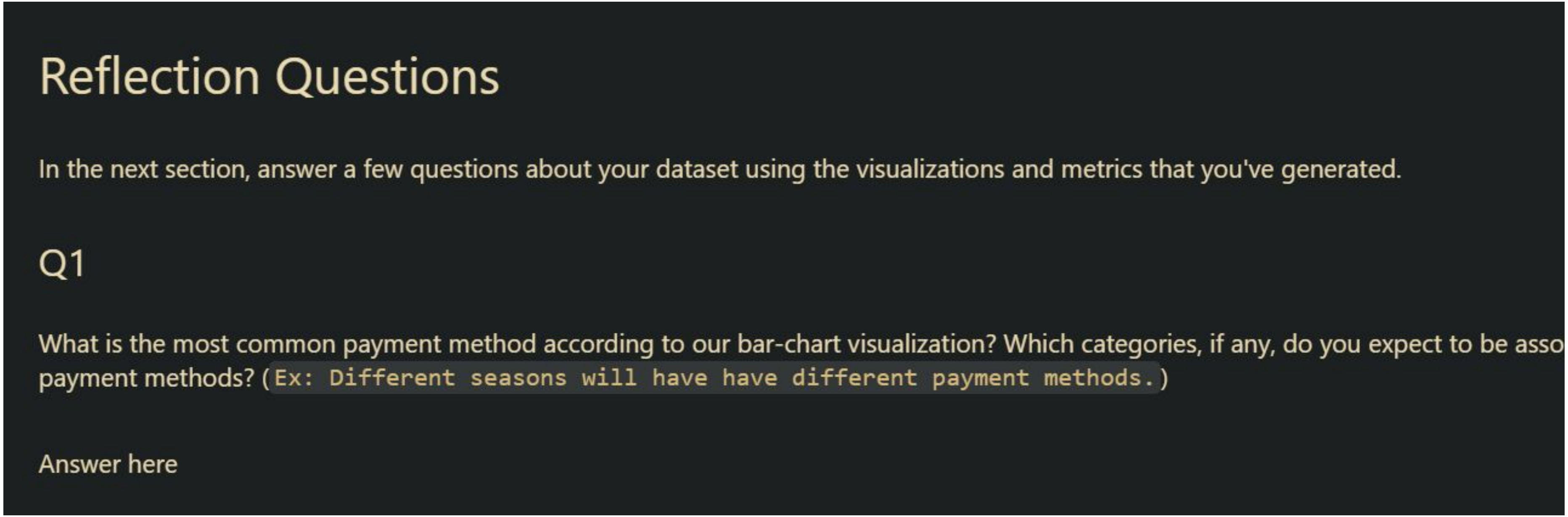
You are a Data Analyst for *FlastFash*, a Budapest-based online clothing store that's looking to break into the American market.

Shopping Dataset Case Study

Today's class will be a highly interactive code-along. For the first half of class we will **work together** (with the help of the wheel) to complete blocks of code in our shopping dataset exploratory analysis.

Some methods will require us to use our research skills to find **documentation on new methods**. After break, we will ask you to complete this case study in your groups.

We will congregate back at 9:20 to discuss results (with the wheels help).



Reflection Questions

In the next section, answer a few questions about your dataset using the visualizations and metrics that you've generated.

Q1

What is the most common payment method according to our bar-chart visualization? Which categories, if any, do you expect to be asso
payment methods? (`Ex: Different seasons will have have different payment methods.`)

Answer here

Prepare a report for your manager by answering the listed reflection questions!

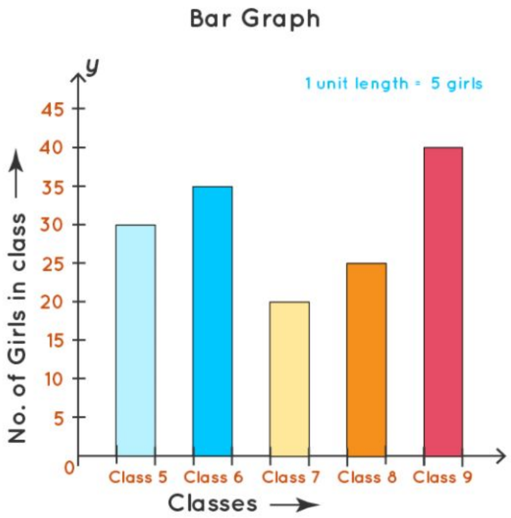
Visualizing Data - Seaborn

Visualizing Data - Bar Graph

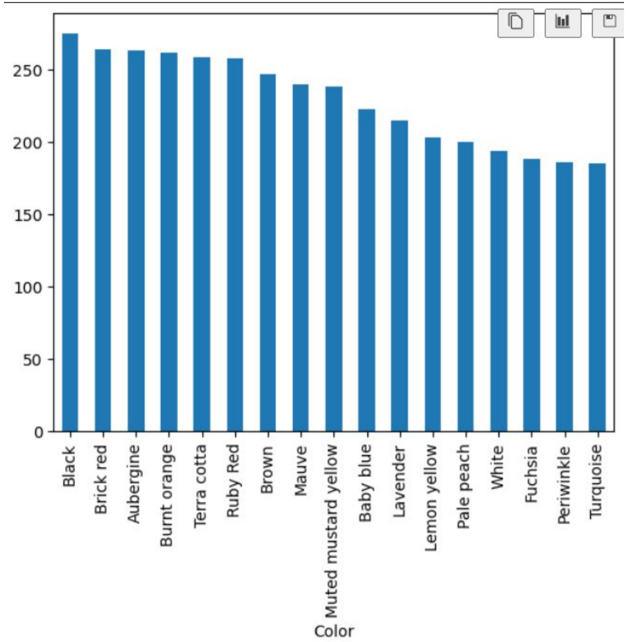
We use **bar-graphs** to represent differences in **categories in one dimension** and sometimes **time**. This visualization is **univariate.**

That is, our **x-axis is always categorical**.

And our **y-axis is always quantitative**.

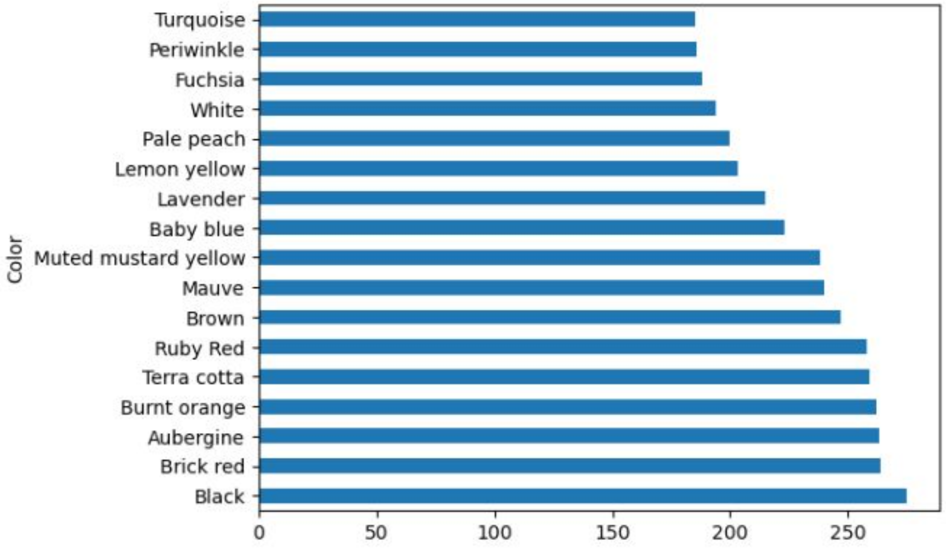


Notice that these x-axis labels aren't easy to read…



df.value_counts("Color").plot.bar()

We can quickly plot the frequencies of categories by specifying a **categorical column** in the **value_counts** method, and then by calling the **plot.bar**() method.



df.value_counts(“Color”).plot.barh()

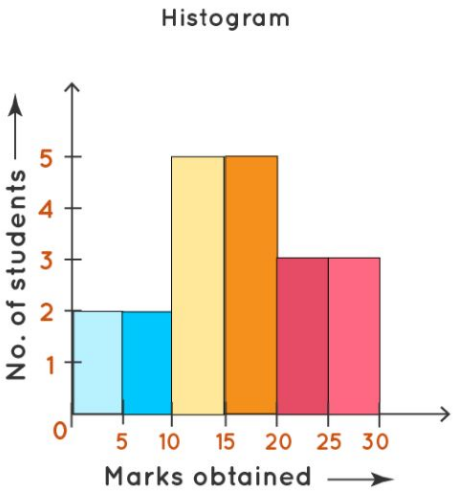
Often when we have a lot of categories, it helps to ease cognitive complexity by creating the **barh**() method to create a horizontal bar plot.

Visualizing Data - Histogram

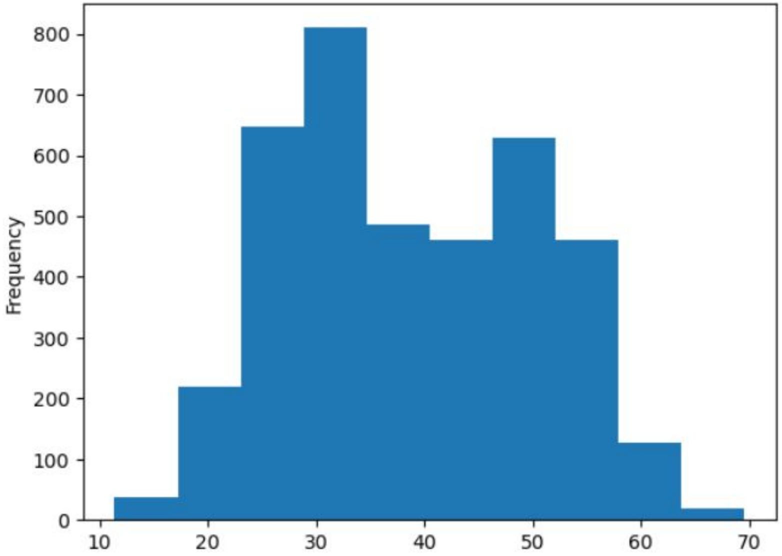
We use **histograms** to represent **distributions of one dimension** (aka **the frequency of different values in a dimension**). This visualization is **univariate.**

That is, our **x-axis is always quantitative**.

And our **y-axis is always quantitative**.

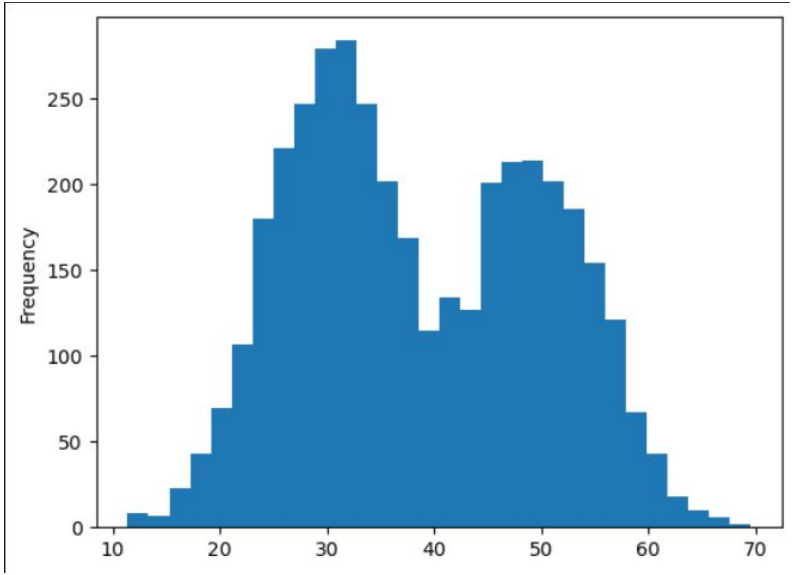


What can we do to better observe our distributions?



df[“Purchase Amount (USD)”].plot.hist()

By specifying a **numerical column** and then calling the **plot.hist()** method, we can plot a histogram on a numeric series to observe the distribution of our dataset.



df[“Purchase Amount (USD)”].plot.hist(bins=30)

By **increasing the number of bins**, we can better observe the distributions that are apparent in our dataset. What kind of distribution do we see here?

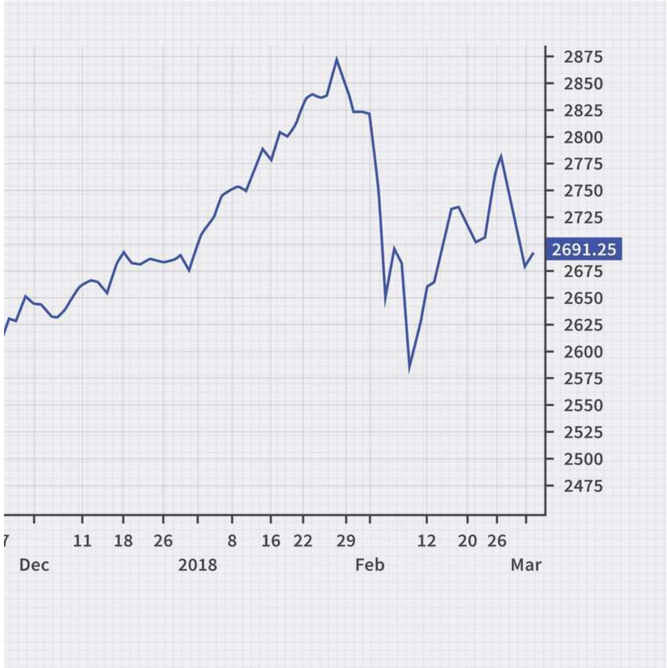
We will explore this type of data visualization tomorrow.

Visualizing Data - Line Plot

We use **line plots** to represent **changes in quantity of one dimension across time.** This visualization is **univariate.**

That is, our **x-axis is always time**.

And our **y-axis is always quantitative**.

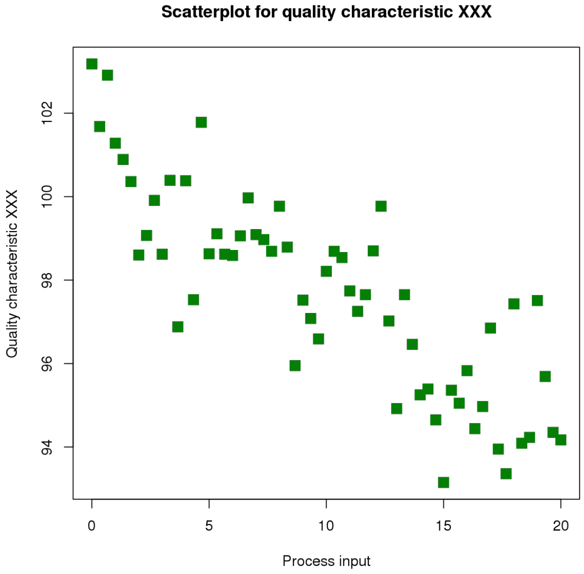


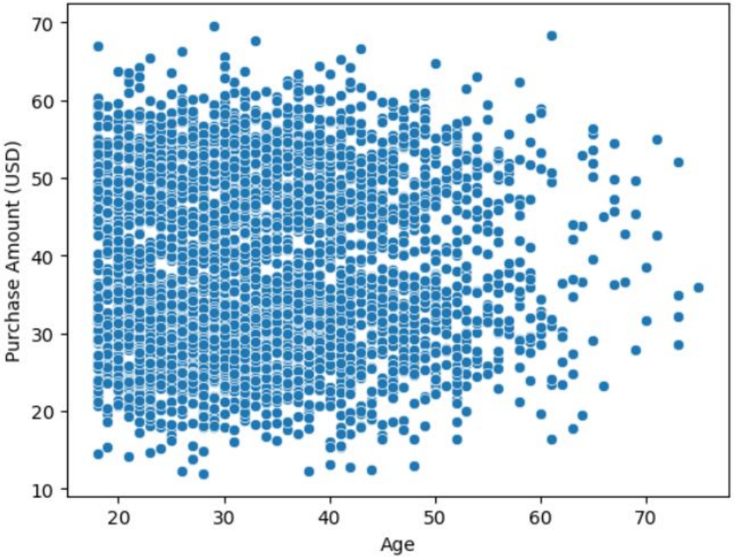
Visualizing Data - Scatter Plot

We use **scatter plots** to represent **distributions of more than one dimensions**. This visualization is **bivariate/multivariate.**

That is, our **x-axis is always quantitative**.

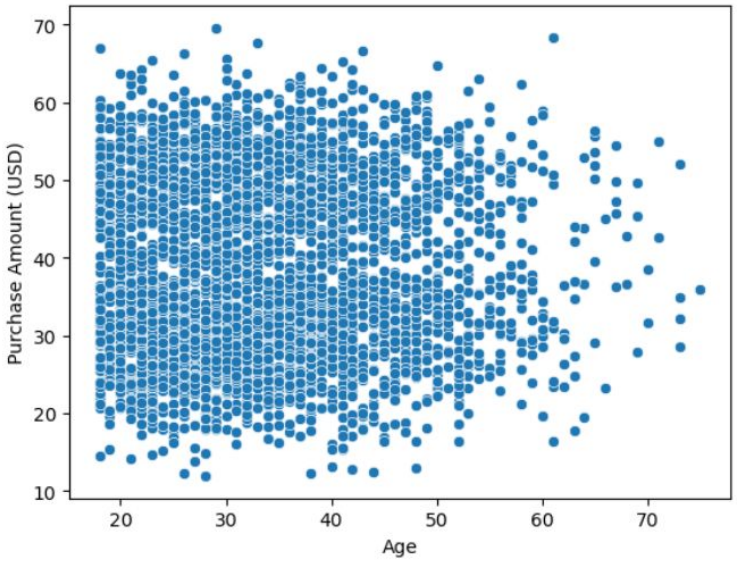
And our **y-axis is always quantitative**.

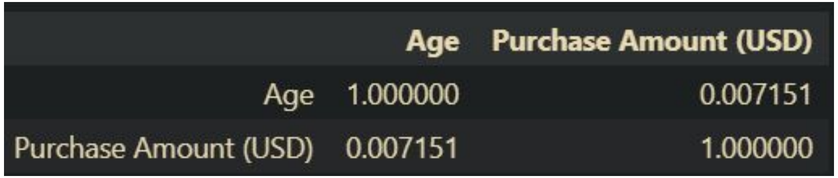




sns.scatterplot(df, x="Age", y="Purchase Amount (USD)")

By specifying a **dataframe, and two numerical columns** in the **scatterplot** method, we can plot a scatter-plot to observe the relationship between two numeric variables. **Do you notice any correlation between age and purchase amount?**





	Age	Purchase Amount (USD)
Age	1.000000	0.007151
Purchase Amount (USD)	0.007151	1.000000

```
df[["Age", "Purchase Amount (USD)"]].corr()
```

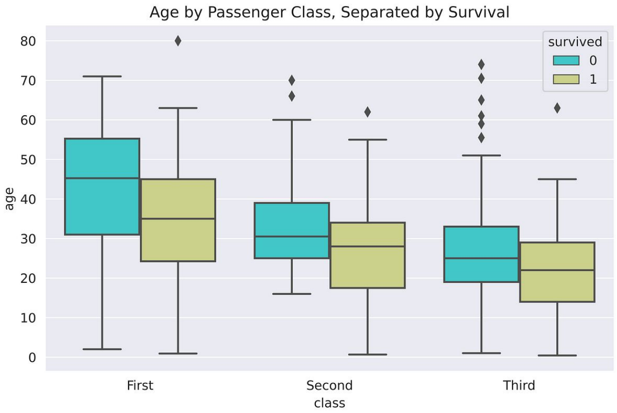
As we see from the correlation matrix, there is no correlation between age and purchase amount. However can you identify **clusters** of data which might be emerging between these two variables?

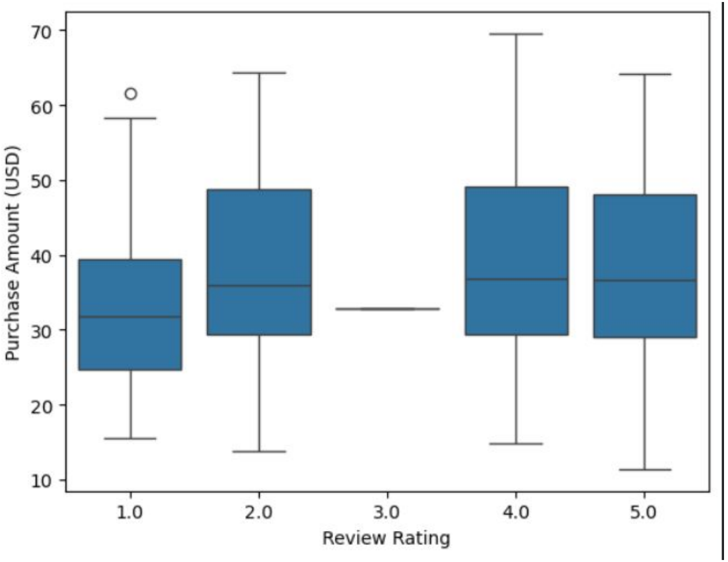
Visualizing Data - Box Plot

We use **box plots** to represent **distributions of different categories in more than one dimension.** This visualization is **bivariate/multivariate.**

That is, our **x-axis is always categorical**.

And our **y-axis is always quantitative**.

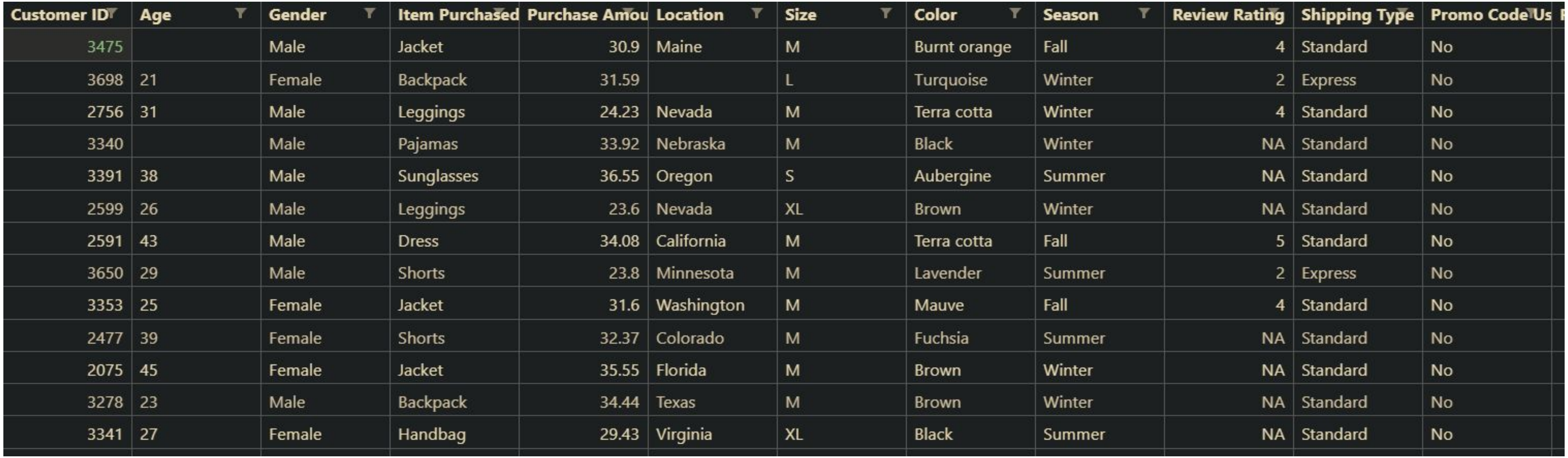




sns.boxplot(df, x="Review Rating", y="Purchase Amount (USD)")

By specifying a **dataframe, one categorical, and one numerical column** in the **boxplot** method, we can plot a box-plot to observe how a distribution varies across categories. **Do you notice any sizeable differences in median?**

Shopping Dataset Case Study



Customer ID	Age	Gender	Item Purchased	Purchase Amou	Location	Size	Color	Season	Review Rating	Shipping Type	Promo Code Us
3475		Male	Jacket	30.9	Maine	M	Burnt orange	Fall	4	Standard	No
3698	21	Female	Backpack	31.59		L	Turquoise	Winter	2	Express	No
2756	31	Male	Leggings	24.23	Nevada	M	Terra cotta	Winter	4	Standard	No
3340		Male	Pajamas	33.92	Nebraska	M	Black	Winter	NA	Standard	No
3391	38	Male	Sunglasses	36.55	Oregon	S	Aubergine	Summer	NA	Standard	No
2599	26	Male	Leggings	23.6	Nevada	XL	Brown	Winter	NA	Standard	No
2591	43	Male	Dress	34.08	California	M	Terra cotta	Fall	5	Standard	No
3650	29	Male	Shorts	23.8	Minnesota	M	Lavender	Summer	2	Express	No
3353	25	Female	Jacket	31.6	Washington	M	Mauve	Fall	4	Standard	No
2477	39	Female	Shorts	32.37	Colorado	M	Fuchsia	Summer	NA	Standard	No
2075	45	Female	Jacket	35.55	Florida	M	Brown	Winter	NA	Standard	No
3278	23	Male	Backpack	34.44	Texas	M	Brown	Winter	NA	Standard	No
3341	27	Female	Handbag	29.43	Virginia	XL	Black	Summer	NA	Standard	No

Complete this analysis and meet back at 9:20 to answer analytical questions via the wheel.

TLAB #3

Doing your Own EDA

DOs for Collaboration

- Discuss **trends & distributions** you notice in your EDA
- Discuss helpful **workflows** & coding concepts
- Point out **resources** your peers can use (*notes, recordings, documentation*)

Don'ts for Collaboration

- **Copy and paste code** from each other
- **Copy and paste code** from ChatGPT

During this grading period, we will be particularly on the lookout for duplicate EDA

Lab (Due 5/14)

Please ensure that your EDA is completed **independently**.

Minas Gerais, Brazil

You are a data engineer at a Brazil-based weather prediction startup called Curu-Sight. The goal of this startup is to analyze weather trends in Brazil and predict the output of non-durable consumer goods at harvest time.

You will analyze a dataset that contains averages calculated based on rainfall, temperature, humidity, and wind metrics collected during the coffee growing season.

You will also analyze a dataset that contains Minas Gerais' crop output. **You will then combine these two datasets to explore how the weather influences coffee growth.**

Tuesday

Tuesday will entail:

- Analysis on a twitter dataset
- Time series analysis
- ...and regex