Data Science I

Overall Points: 63 / 100

Exercise 5: Data Visualization & Machine Learning Intro

Submission Deadline: December 15 2025, 07:00 UTC
University of Oldenburg
Winter 2025/2026
Instructors: Jannik Schröder, Wolfram "Wolle" Wingerath

Submitted by: <Cansu Horata and Charleen Owiti >

Part 1: Exploratory Data Analysis

10 / 25

1.) Provide answers to the questions associated with the following data sets, available at http://www.data-manual.com/data.

a) Analyze the movie data set. What is the range of movie gross in the United States? Which type of movies are most likely to succeed in the market? Comedy? PG-13? Drama? Why?

Solution:

The large range is uneven since there are movies that flop and a few that become Blockbusters and the US gross range from a few thousand dollars to million dollars. Movies that are rated PG tend to do well since they cater to a wider market compared to movies with restriction. In terms of genre, again, Dramas have a narrower audience compared to genres such as comedy which appeals to a greater audience

Did you look at the data ?

b) Analyze the Manhattan rolling sales data set. Where in Manhattan is themost/least expensive real estate located? What is the relationship between sales price and gross square feet?

Solution:

Locations towards the southern parts or Manhattan i.e areas like Upper East side, Soho are insanely expensive whereas areas towards the north like Harlem are less expensive. The relationship between price of properties and gross squarefeet is: typically, the bigger the squarefeet, the more costly the property but this relationship is not linear because some properties with smaller squarefeet in expensive neighbourhood have higher prices compared to properties with bigger squarefeet in less desireable areas.

c) Analyze the 2012 Olympic data set. What can you say about the relationship between a country's population and the number of medals it wins? What can you say about the relationship between the ratio of female and male counts and the GDP of that country?

Solution:

Countries with large populations tend to win more gold medals because of the effect of a large scale which increases chances. But this relationship is again non linear since there are countries with lesser populations bu win several gold medals too.

Countries with higher GDP shows inclusion in supporting both male and female athletes in the olympics. In the data set, it is observeable that the USA had surprisingly more female athletes that male athletes. This is because wealthy countries have more money to invest equally to both male and female sports as compared ti less wealthy countries who just have enough to invest in the men.

d) Analyze the GDP per capita data set. How do countries from Europe, Asia, and Africa compare in the rates of growth in GDP? When have countries faced substantial changes in GDP, and what historical events were likely most responsible for it?

Solution:

We notice from the dataset that increase in GDP or the lack thereof is as a result of certain events that have taken place in the last years such as industralization or technology development or the lack of them. We notice that the GDP of most European countries increases steadily and with stability. Like in the case of France, events such as the renaissance and industrialization and the World wars have affected the GDP growth while those of most African countries like Gabon is volatile with ups and downs. Asian countries like Japan show a rapid increase in GDP.

without

e) Analyze the following data set on electricity demand: https://www.kaggle.com/datasets/albertovidalrod/electricity-consumption-uk-20092022 . Is energy demand different on national holidays compared to normal weekdays or weekends? Can you identify a general trend?

Solution:

could not see the holiday information hence just skipped the task altogether. :(

2.) For one data set of your own choosing, answer the following basic questions:

a) Who constructed it, when, and why?

Solution:

The dataset is from Spotify's public streaming information and published on Kaggle. It covers global music data from 2009 to 2025. The purpose of the dataset is to enable analysis of music trends, artist popularity, and audio characteristics over time.

b) How big is it?

Solution:

The dataset has several row of thousands of songs and each row representing a single song released on Spotify. It includes about 14 columns that describe the song, such as artist name, album name, song duration etc.

c) Identify a few familiar or interpretable records.

Solution:

Many records in the dataset are easily interpretable. For example, artist with many followers Like Taylor swift tend to have high popularity scores and high streaming counts like in her song recent song Fate of Ophelia

d) Find out and describe what Tukey's five number summary is and then provide one for at least 3 different columns.

Solution:

Tukey´s five number summary is a way to describe a dataset's distribution using five key values: Minimum: The smallest data point in the set. First Quartile (Q1): The value below which 25% of the data falls (25th percentile). Median (Q2): The middle value; 50% of data is below, 50% is above (50th percentile). Third Quartile (Q3): The value below which 75% of the data falls (75th percentile). Maximum: The largest data point in the set.

Applied to the Spotify dataset, the popularity column shows a right-skewed distribution, with a small number of highly popular songs and many tracks with low popularity and the artist number of followers shows a highly skewed distribution, highlighting the unequal concentration of audience attention among artists.

When it comes to song duration, the minimum duration corresponds to very short interludes or skits, while the maximum reflects unusually long tracks. The median duration lies within a relatively narrow range, indicating that most songs have similar lengths. The presence of long-duration outliers suggests occasional experimental or extended recordings.

Number of songs per album: The minimum represents singles or very short albums, while the maximum captures albums with unusually large tracklists. The interquartile range shows that most albums contain a moderate number of songs, with a few outliers representing compilations or deluxe editions

e) State at least one interesting or noteworthy thing that you Learned from your data set.

Solution:

One observation from the dataset is the strong concentration of popularity around a small number of artists. For example, artists such as Taylor Swift appear repeatedly among highly popular tracks and albums, often with consistently high popularity

values across multiple releases. This highlights how global streaming platforms tend to amplify already successful artists, leading to a highly unequal distribution of attention where a few artists dominate listening behavior over long periods of time.

Part 2: Interpreting Visualizations

3 / 25

3.) Search your favorite news websites until you find 4 interesting charts/plots, ideally half good and half bad. For each, please critique along the following dimensions:

a) Does it do a good job or a bad job of presenting the data? Why?
b) Does the presentation appear to be biased, either deliberately or accidentally?
c) Is there chartjunk in the figure? Where?
d) Are the axes labeled in a clear and informative way?
e) Is the color used effectively?
f) How can the graphic be improved?

https://www.dw.com/en/africa-at-the-center-of-uschina-resource-race/a-75130071
a) The chart does a good job of showing the geographic distribution of US and Chinese resource involvement in Africa. The map format is appropriate for highlighting spatial patterns.

3 missing

b) There is no clear deliberate bias, but the visualization may unintentionally emphasize rivalry by framing the issue mainly as a US–China competition.

c) Chartjunk is limited, though multiple icons and labels slightly clutter regions with many investments.

d) As a map-based chart, there are no traditional axes, which reduces quantitative precision and makes exact comparisons difficult.

e) Color is used effectively to distinguish between US and Chinese involvement, improving readability.

f) The chart could be improved by adding quantitative information (e.g., investment size) or a supplementary bar chart.

4.) Visit https://viz.wtf and find five laughably bad visualizations. Explain why they are both bad and amusing.

1.Pie chart totaling more than 100% This visualization is bad because pie charts represent parts of a whole, and exceeding 100% violates basic logic. It is amusing because it confidently presents mathematically impossible information.

without images or links i cant evaluate this

2.3D exploding pie chart The 3D perspective distorts slice sizes, making accurate comparison impossible. It is funny because the dramatic explosion adds visual noise without adding information.

3.Dual y-axis chart with unrelated variables This chart falsely suggests a relationship between two unrelated variables by placing them on separate axes. It is amusing because the apparent correlation is entirely artificial.

4.Overloaded word cloud used as data analysis Word clouds lack numerical precision and make comparison difficult. It is laughable because it looks analytical while conveying very little actual information.

5.Rainbow color scale without a legend Using many colors without explanation makes the chart impossible to interpret. The humor comes from the visual chaos and confusion it creates.

Part 3: Creating Visualizations & Storytellin 25 / 25

5.) Construct a revealing visualization of some aspect of your favorite data set, using:

a) A well-designed table.

In [3]:
```
import pandas as pd

df = pd.read_csv("shopping_behavior_updated.csv")

table = (
    df.groupby(["Category", "Gender"])["Purchase Amount (USD)"]
    .mean()
    .reset_index()
)

table
```

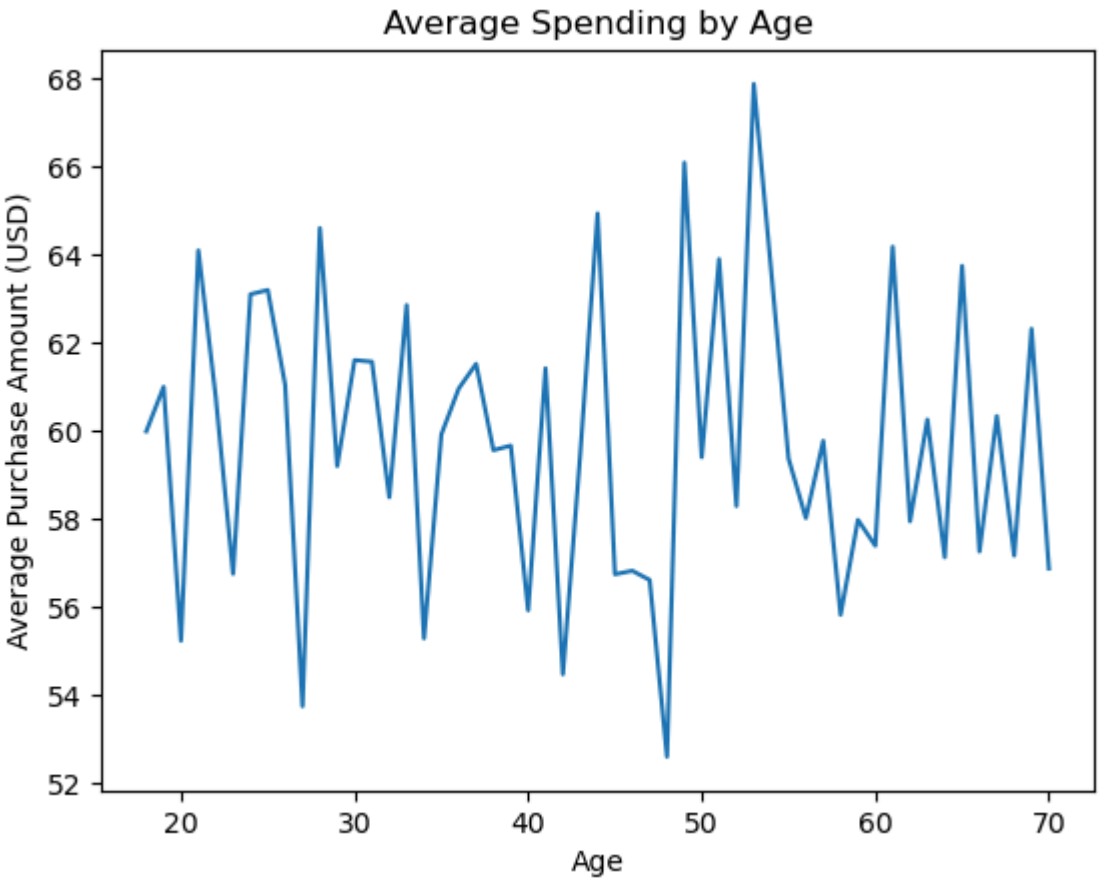
Out[3]:

	Category	Gender	Purchase Amount (USD)
0	Accessories	Female	60.762755
1	Accessories	Male	59.411557
2	Clothing	Female	60.496403
3	Clothing	Male	59.803556
4	Footwear	Female	59.472362
5	Footwear	Male	60.645000
6	Outerwear	Female	58.425743
7	Outerwear	Male	56.605381

Hier Text eingeben

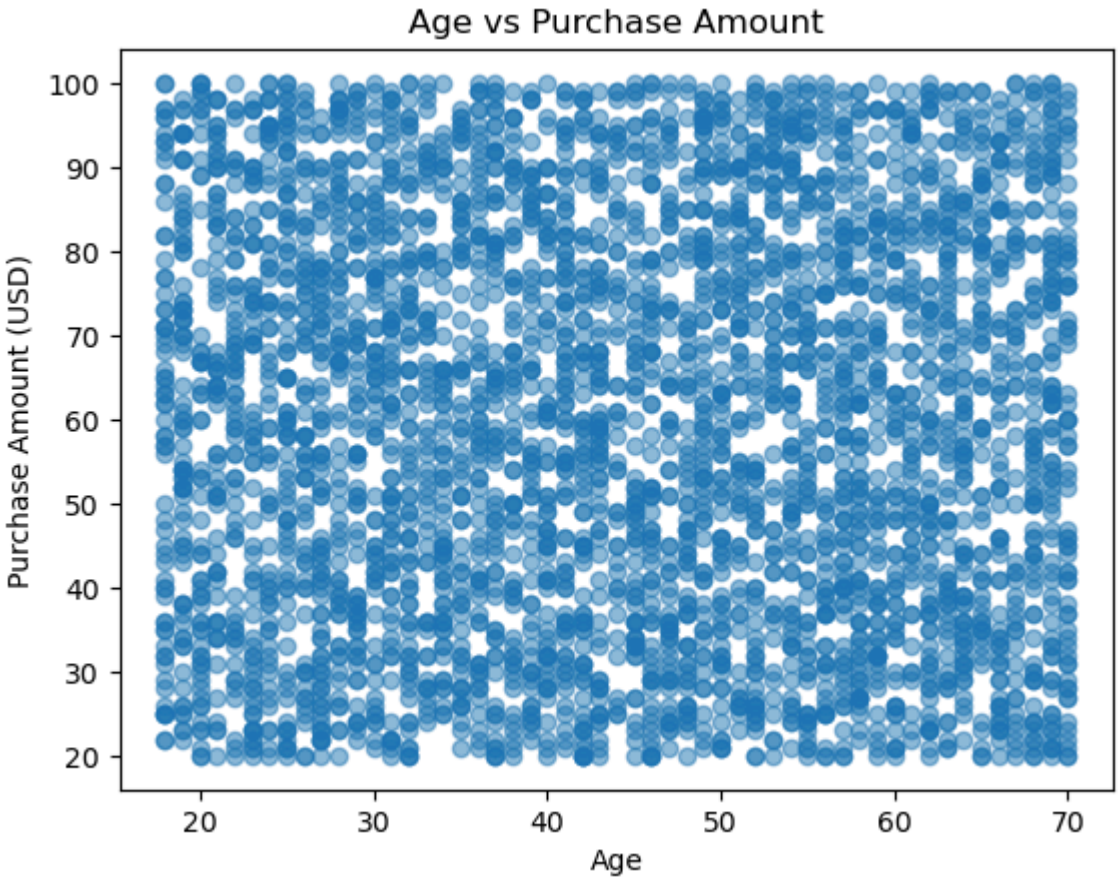
b) A dot and/or line plot.

In [5]:
```
import matplotlib.pyplot as plt

age_spending = df.groupby("Age")["Purchase Amount (USD)"].mean()

plt.figure()
plt.plot(age_spending.index, age_spending.values)
plt.xlabel("Age")
plt.ylabel("Average Purchase Amount (USD)")
plt.title("Average Spending by Age")
plt.show()
```



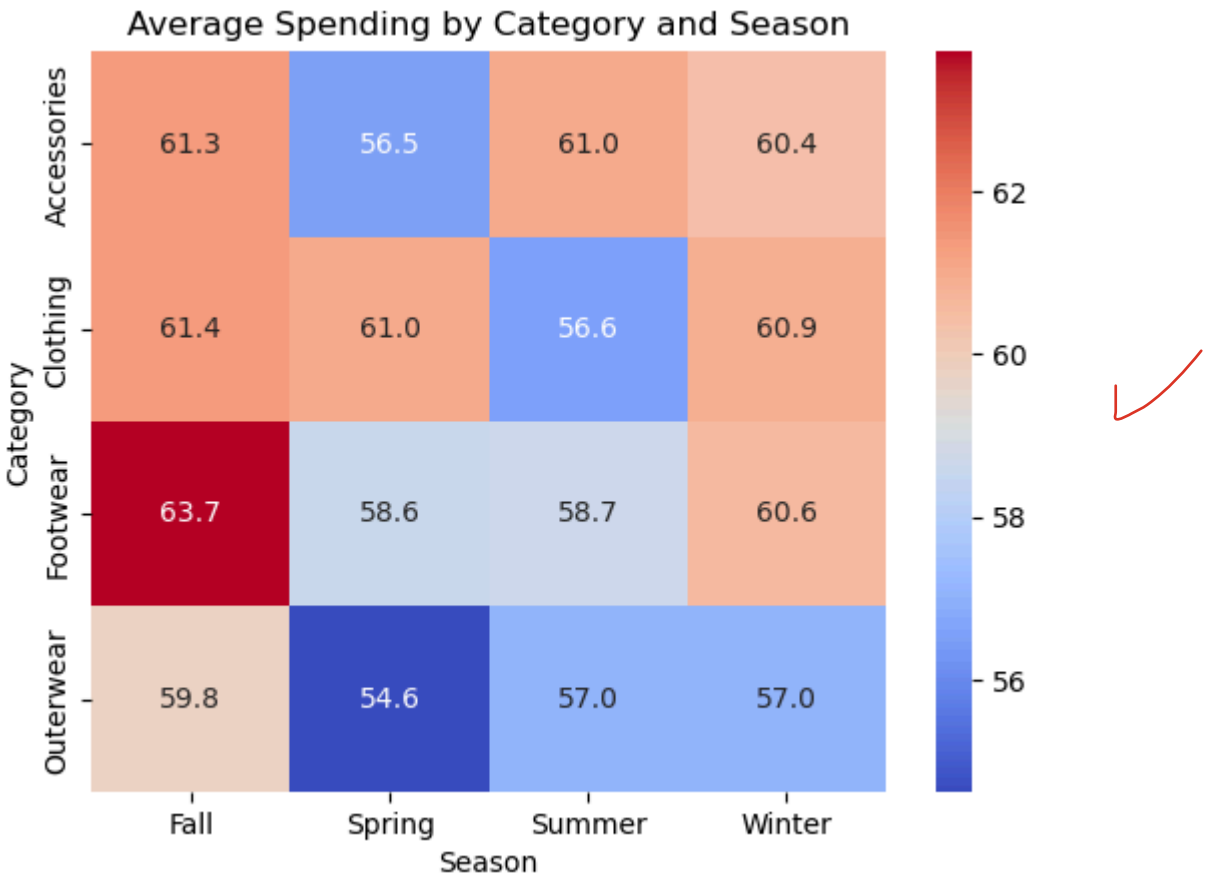
c) A scatter plot.

In [6]:
```
plt.figure()
plt.scatter(df["Age"], df["Purchase Amount (USD)"], alpha=0.5)
plt.xlabel("Age")
plt.ylabel("Purchase Amount (USD)")
plt.title("Age vs Purchase Amount")
plt.show()
```



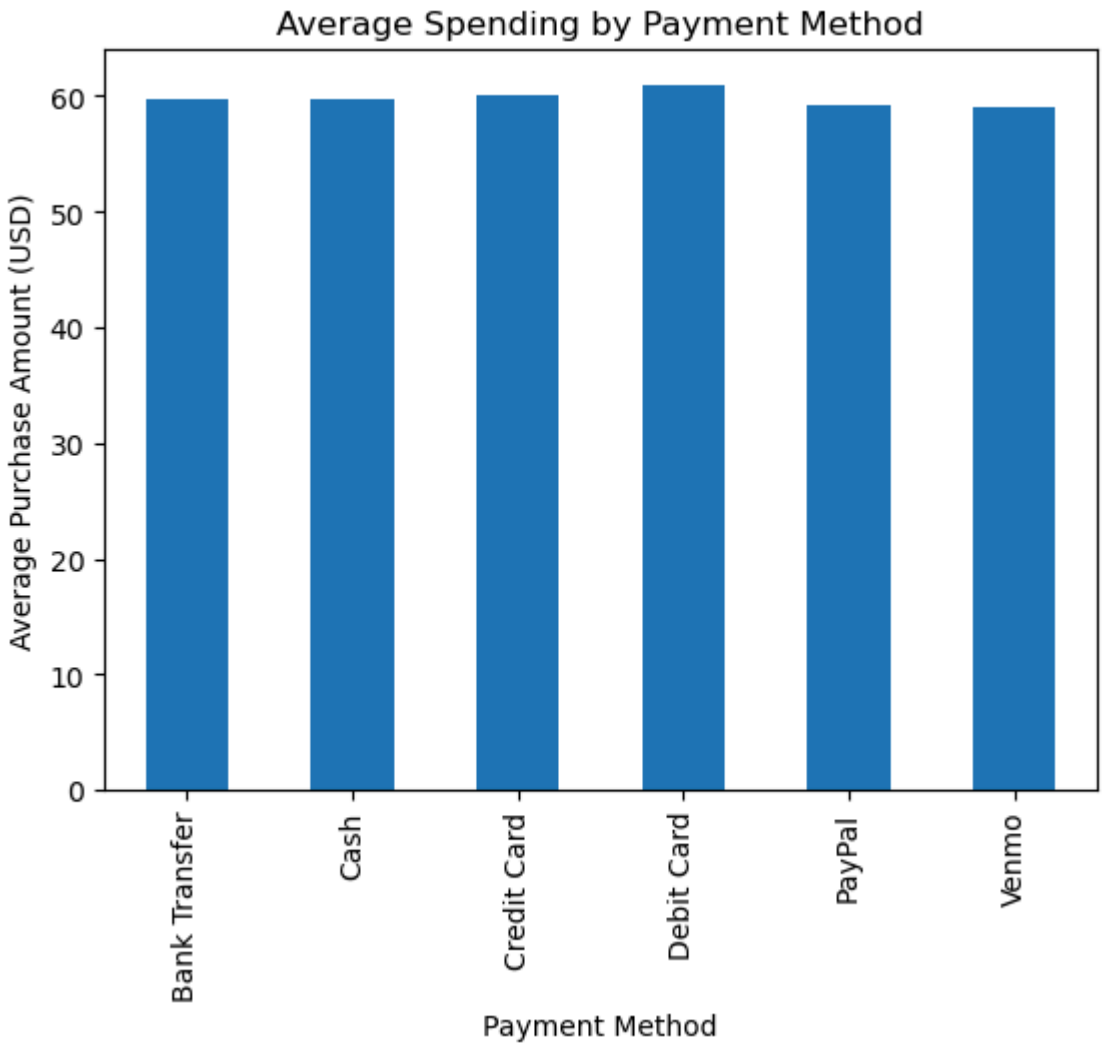
d) A heatmap.

In [8]:
```
import seaborn as sns
import matplotlib.pyplot as plt

pivot = df.pivot_table(
    values="Purchase Amount (USD)",
    index="Category",
    columns="Season",
    aggfunc="mean"
)

sns.heatmap(pivot, annot=True, fmt=".1f", cmap="coolwarm")
plt.title("Average Spending by Category and Season")
plt.show()
```



e) A bar plot and/or a pie chart.

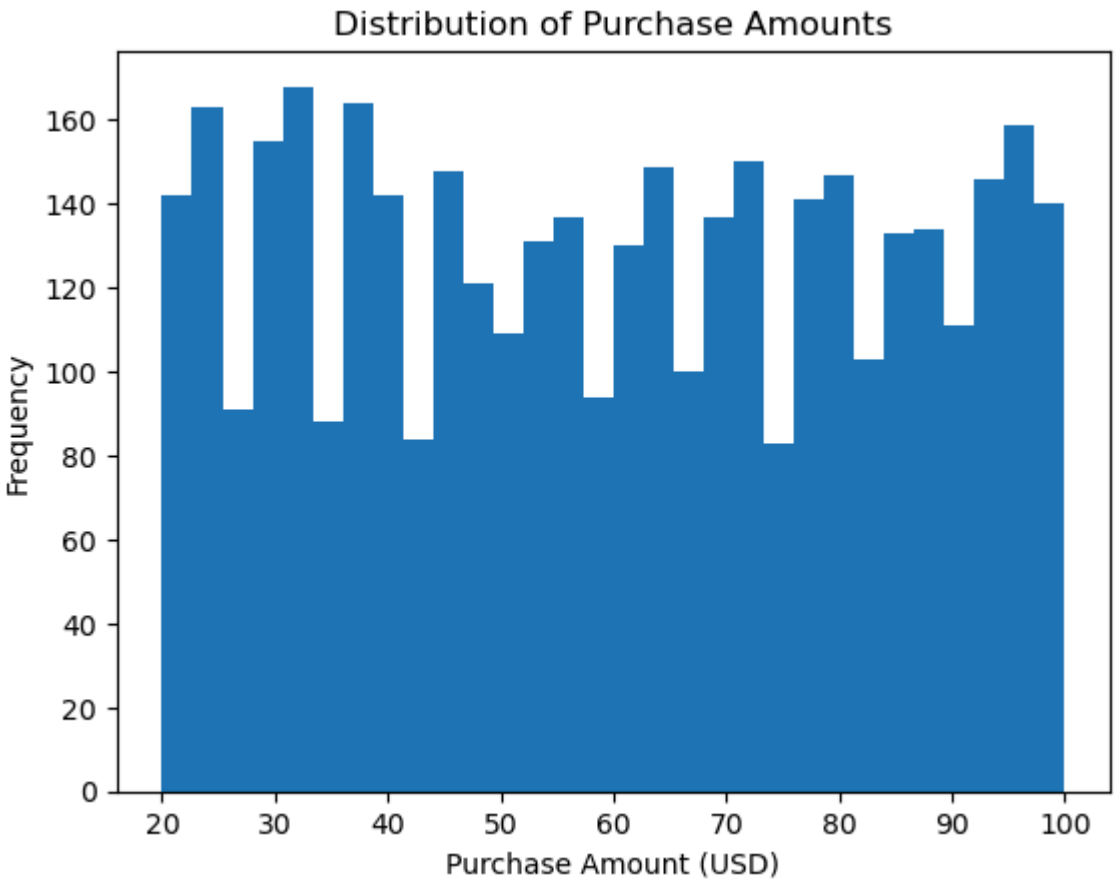
In [9]:
```
payment_avg = df.groupby("Payment Method")["Purchase Amount (USD)"].mean

payment_avg.plot(kind="bar")
plt.ylabel("Average Purchase Amount (USD)")
plt.title("Average Spending by Payment Method")
plt.show()
```



f) A histogram.

In [10…

```
plt.figure()
plt.hist(df["Purchase Amount (USD)"], bins=30)
plt.xlabel("Purchase Amount (USD)")
plt.ylabel("Frequency")
plt.title("Distribution of Purchase Amounts")
plt.show()
```



6.) Find and tell a Story with data! To this end, first go to https://uol.de/planung-entwicklung/akademisches-controlling/studium-und-lehre
and select one of the following data sets:

- **Studienanfänger*innen / Studienanfänger*innen (Fallstatistik) nach Studiengang**
- **Fachstudiendauer / Übersicht über die Fachstudiendauer**

Then explore the data and find a story to tell with it.

a) Define an audience and a goal.
(Example: A data viz that highlights a potential issue to the university council or one that tries to win new students for a certain subject area.)

In [30…

```
import pandas as pd

# 1) Remove unnamed columns and strip names
df = df.loc[:, ~df.columns.astype(str).str.contains("^Unnamed")]
df.columns = df.columns.astype(str).str.strip()

# 2) Make duplicate column names unique (very common in this Excel)
# Corrected approach to handle duplicate column names
new_columns = []
seen = {}
for col in df.columns:
    if col in seen:
        seen[col] += 1
        new_columns.append(f"{col}_{seen[col]}")
```

```
    else:
        seen[col] = 0
        new_columns.append(col)
df.columns = new_columns

# 3) Forward fill merged-cell columns
for c in ["Fakultät", "Lehreinheit", "Studienfach", "Abschluss"]:
    if c in df.columns:
        df[c] = df[c].ffill()

# 4) Convert numeric columns safely
numeric_cols = [c for c in df.columns if str(c).isdigit() or c in ["gesa

for c in numeric_cols:
    # convert to string, replace German commas, then numeric
    s = df[c].astype("string").str.replace(",", ".", regex=False)
    df[c] = pd.to_numeric(s, errors="coerce")

df.head(10)
```

Out[30]:

	nan	nan_1	nan_2	nan_3	nan_4	201
13	I	Pädagogik	Bildungs/Wissenschaftsman	Fach-Master	5	8.09090
14	I	Pädagogik	Erzieh.-Bildungswissensch	Fach-Master	4	5.64705
15	I	Pädagogik	Interk.-Bildung/Beratung	Fach-Bachelor	6	11.66666
16	I	Pädagogik	Päd. Hand. Migrationsges.	Fach-Bachelor	4	<NA
17	I	Pädagogik	Pädagogik	Fach-Bachelor	6	6.69230
18	I	Pädagogik	Pädagogik	Zwei-Fächer-Bachelor	6	6.97435
19	I	Sachunterricht	Interdisz. Sachbildung	Zwei-Fächer-Bachelor	6	6.45833
20	I	Sachunterricht	Sachunterricht	Master Ed. Grundschule	4	<NA
21	I	Sachunterricht	Sachunterricht	Master Ed. Sonderpäd	4	4.66666
22	I	Sachunterricht	Sachunterricht	Master Ed. Gr.-/Hauptsch.	2	3.72222

In [32…

```
df.columns = [
    "Fakultät",
    "Lehreinheit",
    "Studienfach",
    "Abschluss",
    "Regelstudienzeit",
    "2015",
    "2016",
    "2017",
    "2018",
    "2019",
    "2020",
    "2021",
    "2022",
    "2023",
```

```
    "2024",
    "gesamt",
    "Anzahl Abschlüsse"
]
df.head(10)
```

Out[32]:

	Fakultät	Lehreinheit	Studienfach	Abschluss	Regelstudie
13	I	Pädagogik	Bildungs/Wissenschaftsman	Fach-Master	
14	I	Pädagogik	Erzieh.-Bildungswissensch	Fach-Master	
15	I	Pädagogik	Interk.-Bildung/Beratung	Fach-Bachelor	
16	I	Pädagogik	Päd. Hand. Migrationsges.	Fach-Bachelor	
17	I	Pädagogik	Pädagogik	Fach-Bachelor	
18	I	Pädagogik	Pädagogik	Zwei-Fächer-Bachelor	
19	I	Sachunterricht	Interdisz. Sachbildung	Zwei-Fächer-Bachelor	
20	I	Sachunterricht	Sachunterricht	Master Ed. Grundschule	
21	I	Sachunterricht	Sachunterricht	Master Ed. Sonderpäd	
22	I	Sachunterricht	Sachunterricht	Master Ed. Gr.-/Hauptsch.	

Audience: University management and program coordinators at the University of Oldenburg.

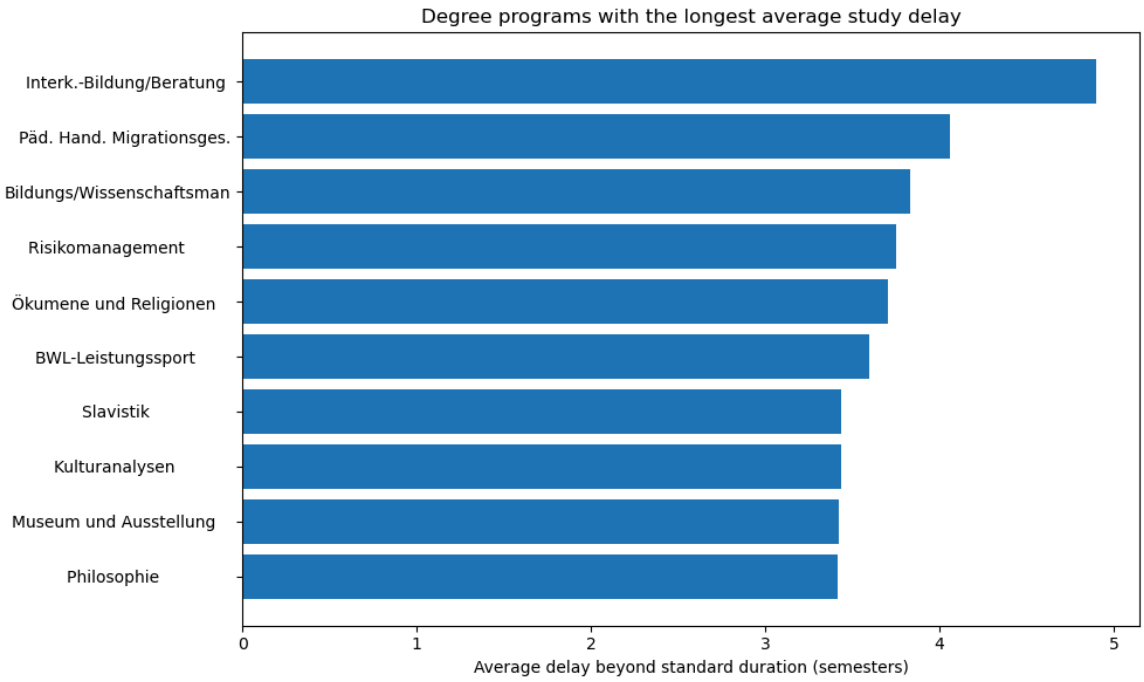
Goal: The goal is to use official data on average study duration to identify degree programs in which students systematically take longer than the standard study duration. This information can support data-driven decisions regarding academic advising, curriculum design, and resource allocation in order to improve study progression and timely graduation.

b) Create a (communicative) data visualization to help your cause.

In [33…

```
import matplotlib.pyplot as plt

# Calculate delay (actual duration - standard duration)
df["Delay"] = df["gesamt"] - df["Regelstudienzeit"]

# Focus on reliable programs (enough graduates)
plot_data = (
    df[df["Anzahl Abschlüsse"] >= 30]
    .sort_values("Delay", ascending=False)
    .head(10)
)

# Plot
plt.figure(figsize=(10,6))
plt.barh(plot_data["Studienfach"], plot_data["Delay"])
```

```
plt.gca().invert_yaxis()
plt.xlabel("Average delay beyond standard duration (semesters)")
plt.title("Degree programs with the longest average study delay")
plt.tight_layout()
plt.show()
```



Part 4: Machine Learning Intro

25 / 25

7.) Give decision trees to represent the following Boolean functions:

a) A and B.

Solution:

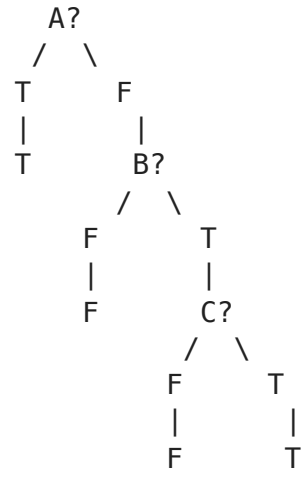
Output is true if and only if both A and B are true. So, if A is true, we have to check if B is true or false but if A is false, no need to check B since the output is already false.

A¿ /
F T | | F B¿ /
F T | | F T

b) A or (B and C).

Solution:

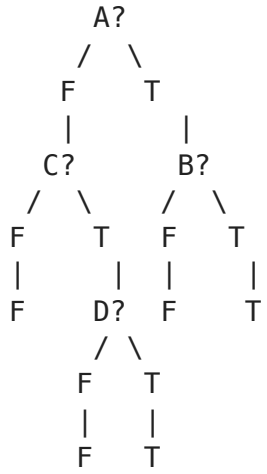
Output is true if A is true or both B and C are true. If A lone is true, then the function is true since it requires either A or the pair, B and C to be true. However, if A is false, then the output can only be true if both B and C are true



c) (A and B) or (C and D)

Solution:

The function is true if both A and B are true OR C and D is true. A is first evaluated and if A is true, then B is evaluated. If A is false, nor need to check B since the pair is already False and C is checked directly. But if A and B both turn out to be true, then the output is true and no need to evaluate C or B.



8.) Consider the following titanic dataset:
https://www.kaggle.com/competitions/titanic/data

a) Load the test and training data sets. Briefly describe the dataset.

Solution:

The dataset is a about Passangers aboard the Royal Mail Ship Titanic with columns such as relations to show whether they had siblings and or Parent, age of passangers, binanry of whether they survived or not among other columns. There is another test dataset that excludes the binanry column about the survival as this is left out for predictability.

b) Train a random forest classifier to predict survival chances for Titanic passengers.
(Hint: You can use one of the tutorials/submissions as a starting point.)

Solution:

In [16…

```
# First, we define the missing variables

import pandas as pd
import numpy as np
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split

# Load the data
train_df = pd.read_csv("/Users/charleen/Desktop/train.csv")
test_df = pd.read_csv("/Users/charleen/Desktop/train.csv")

# Prepare attributes and target
X_train = train_df.drop(['Survived', 'PassengerId', 'Name', 'Ticket', 'C
X_train = pd.get_dummies(X_train)  # Convert categorical variables
y_train = train_df['Survived']

# Prepare test data
X_test = test_df.drop(['PassengerId', 'Name', 'Ticket', 'Cabin'], axis=1
X_test = pd.get_dummies(X_test)

X_test = X_test.reindex(columns=X_train.columns, fill_value=0)

rf_model_full = RandomForestClassifier(
    n_estimators=100,
    max_depth=10,
    min_samples_split=5,
    min_samples_leaf=2,
    max_features='sqrt',
    random_state=42,
    n_jobs=-1,
    class_weight='balanced'
)


rf_model_full.fit(X_train, y_train)
```

```
test_predictions = rf_model_full.predict(X_test)

# Create file
submission = pd.DataFrame({
    'PassengerId': test_df['PassengerId'],
    'Survived': test_predictions
})

# Save the said file
submission.to_csv('titanic_predictions.csv', index=False)

print(f"Number of survivors predicted: {test_predictions.sum()} out of {
print(f"Survival rate predicted: {(test_predictions.sum()/len(test_predi
```

```
Number of survivors predicted: 320 out of 891
Survival rate predicted: 35.9%
```

c) Evaluate the performance of your model and iterate on it to improve it!

Solution:

In [17…

```
from sklearn.model_selection import cross_val_score

cv_scores = cross_val_score(rf_model_full, X_train, y_train,
                            cv=5, scoring='accuracy', n_jobs=-1)
print(f"\nCross-Validation Accuracy: {cv_scores.mean():.4f} (+/- {cv_scc
```

```
Cross-Validation Accuracy: 0.8171 (+/- 0.0539)
```

Finally: Submission

Save your notebook and submit it (as both **notebook and PDF file**). And please don't forget to ...

- ... choose a **file name** according to convention (see Exercise Sheet 1, but please **add your group name as a suffix** like `_group01`) and to
- ... include the **execution output** in your submission!