

✓ EXAM FIN 687 Python in Finance

https://colab.research.google.com/drive/1TzmlEx_v7vo8n92GBmwxoWOhlaFobz--

Edit this cell by double clicking it. Then enter your name and student id number:

- Name:
- Student ID:

Submission guidelines

- After you solve the exercises, you have to hand in one ZIP file that contains two files:
 - The exported .ipynb file from Google Colab
 - The "printed" notebook as a PDF
- Check the Outlook slides on Ilias if you need guidance on how to obtain the two files.
- Your notebook has to run in one single go from start to finish
 - Test this yourself by clicking on `Runtime > Restart session and run all` in the menu bar.
- Please call your files
 - FIN687_Exam_<Lastname>_<StudentID>.zip
 - FIN687_Exam_<Lastname>_<StudentID>.ipynb
 - FIN687_Exam_<Lastname>_<StudentID>.pdf
- You have to submit your solutions electronically on the Ilias for examinations (where you downloaded the tasks). If you encounter any technical problems, please notify me immediately at sven.vahlpahl@uni-mannheim.de. If you cannot upload your solutions, please send them to me via email (but only then) before the deadline.

General rules

- Add your name and student ID to the top of the notebook.
- Whether or not you follow good programming practices also affects the points you receive for each question. For example:
 - Structure your code clearly. Use text cells or comments to explain what your code does.
 - Use meaningful variable and file names.
 - Avoid unnecessary loops.
- Do not use any other software to (or manually) process the data besides Python.
- Use only the packages discussed in the lecture for regressions (`pyfixest`) and t-tests (`pingouin`).
- Variable names need to be unique within each problem.
- You have to solve the exam by yourself. You are not allowed to work in groups.
- Some tasks require some minimal interpretation of the results. In these cases, 1 or 2 sentences are enough.
- Whenever we refer to logarithms, we mean the natural logarithm.
- You are not allowed to use AI assistants to write finished code for you.
- Ensure your code runs from top to bottom by using **Restart session and run all**.
- Finally: Don't get overwhelmed. While there are many tasks, most of them can be solved with only a few lines of code. In fact, some of them even only require a single line.

✓ Packages and Imports

```
# Install your packages in this cell
```

```
# Import your packages in this cell
```

✓ Problem 1: Fundamentals [5 points]

```
companies = ['ABC Corp.', 'Meta Platforms', 'Amazon Inc.', 'Tesla Motors']
```

```
numbers = [3.2, 4.5, 8.9, "missing", 6.7, None]
```

Part a [2 points]

Work with the provided variable `companies`.

- Make a new **tuple** that contains only company names that contain the letter `t` (case insensitive).

Part b [3 points]

- Go over each one of the values in the provided variable `numbers` and convert it to an integer. Store the results. If it's not possible, show an error message indicating the current iteration, but keep going through the rest of the values.

✓ Problem 2: Data exploration [25 points]

```
# This dataset is shared with the next problem  
!gdown 1QT1dQYvbfsh0vAGJr1ACIjmI3e-UMYkG
```

You have been provided with the dataset `walmart.csv`, which contains information about weekly sales at walmart, in addition to other datapoints.

Column Name	Description
Store	Store identification number
Date	Sales week start date
Weekly_Sales	Weekly sales in USD
Holiday_Flag	Mark on the presence or absence of a holiday (1/0)
Temperature	Air temperature in the region in Fahrenheit
Fuel_Price	Fuel cost in the region in USD per gallon
CPI	Consumer price index
Unemployment	Unemployment rate (0-100)

Part a [15 points]

- Ensure the date is parsed appropriately.
- Which data types are the columns?
- How many missing values are in each column?

Part b [10 points]

- Plot a seaborn pairplot for the data.
- Export it as a png picture without borders and high resolution.
- What is shown on the diagonal? What is shown on the off-diagonals?
- Look at the weekly_sales - store combination. Why are there these "strings" of points?

✓ Problem 3: Analysis [60 points]

```
# This dataset is shared with the previous problem
!gdown 1QT1dQYvbfsh0vAGJr1ACIjmI3e-UMYkG
```

You have been provided with the dataset `walmart.csv`, which contains information about weekly sales at walmart, in addition to other datapoints.

Column Name	Description
Store	Store identification number
Date	Sales week start date
Weekly_Sales	Weekly sales in USD
Holiday_Flag	Mark on the presence or absence of a holiday (1/0)
Temperature	Air temperature in the region in Fahrenheit
Fuel_Price	Fuel cost in the region in USD per gallon
CPI	Consumer price index
Unemployment	Unemployment rate (0-100)

Please note: **Only use the regression and t-testing packages we used in the lecture!**

Part a [10 points]

- Convert the temperature to degrees Celsius, the fuel price to price per liter, and unemployment to a 0-1 percentage.
- Add last week's sales as a variable.
- Drop rows that have missing values.

Use the newly created variables in your regressions.

Part b [15 points]

- Take a look at the descriptive statistics below this cell labeled `Goal` for part b.

- Reproduce this table. Details matter.

Part c [10 points]

- Visualize the correlations between all numeric variables in a heatmap.
- Test whether the mean of the weekly_sales is significantly different between Decembers and Januaries for the store with identifier 4.

Part d [10 points]

Run regressions to answer the following research questions.

- (Code only) Is temperature related to the weekly sales, while controlling for last week's sales?
 - (Text only) Interpret your results: Is the temperature related to sales?
- (Code only) How do last week's sales, temperature, fuel prices, holidays, and unemployment rates collectively impact current weekly sales?
 - (Text only) Do the results seem intuitive to you? Why or why not.

Part e [15 points]

Run regressions to answer the following research questions (code only). If you transform sales, also transform last week's sales.

1. What is the log-linear relationship between sales, accounting for the effect of previous week's sales, temperature, fuel prices, and unemployment rates?
2. In a log-linear model that accounts for store-specific and date-specific variations, how do previous week's sales, temperature, fuel prices, and unemployment rates relate to current weekly sales?
3. Show your results in a table. Rename the variables in the table to more natural names. Show the t statistic in the line underneath the coefficient.

Answer the following questions based on the model used in 2. (text only):

4. Why can you not include holiday_flag?
5. Give two examples of potential factors we are capturing by controlling for store-specific variation.

Goal for part b

Descriptive Statistics					
Walmart Data Set					
	Sales	Temperature	Fuel Price	Consumer Price Index	Unemployment Rate
mean	\$1,047,072	15.92°C	\$0.89	171.6	8.0%
std	\$564,345	10.25°C	\$0.12	39.4	1.9%
min	\$209,986	-18.92°C	\$0.65	126.1	3.9%
2.5%	\$280,911	-4.88°C	\$0.69	126.1	4.6%
50%	\$960,846	17.04°C	\$0.91	182.6	7.9%
97.5%	\$2,168,051	32.40°C	\$1.09	223.9	14.0%
max	\$3,818,686	37.86°C	\$1.18	227.2	14.3%

✓ Problem 4: Web scraping [30 points]

Part a [20 points]

How many glass baubles are offered by the manufacturer Premier on <https://www.thechristmasshop.co.uk/c/for-the-tree/baubles/glass> ?

Part b [10 points]

Plot a bar chart of the number of products per manufacturer sorted with the highest value on the left.

✓ Problem 5: Machine learning [45 points]

```
!gdown 1033a46bnmKC82jCJ6VV8z8sDVIot5XvK
```

You are provided with the dataset `loans.csv`, which contains information about borrowers and a target variable `loan_status`.

Column Name	Description
loan_id	Unique identifier for each loan application
no_of_dependents	Number of people dependent on the loan applicant
education	Education level of the loan applicant
self_employed	Whether the applicant is self-employed (Yes/No)
income_annum	Annual income of the applicant
loan_amount	Amount of loan requested
loan_term	Duration of the loan in months
credit_score	Credit score of the applicant
residential_assets_value	Value of residential assets owned by applicant
commercial_assets_value	Value of commercial assets owned by applicant
luxury_assets_value	Value of luxury assets owned by applicant
bank_asset_value	Value of assets held in bank accounts
loan_status	Status of the loan application (Approved/Rejected)

Your goal is to build a model that predicts whether the loan should be approved or rejected.

Make sure you set a random state.

Part a [15 points]

- Is this a regression or a classification problem?
- Ensure you can use all features in your analysis.
- Split the data into 75% train, 25% data.

Part b [20 points]

- Select and train an appropriate model. Ensure you are not overfitting.
- Evaluate the model using the `f1_score`.