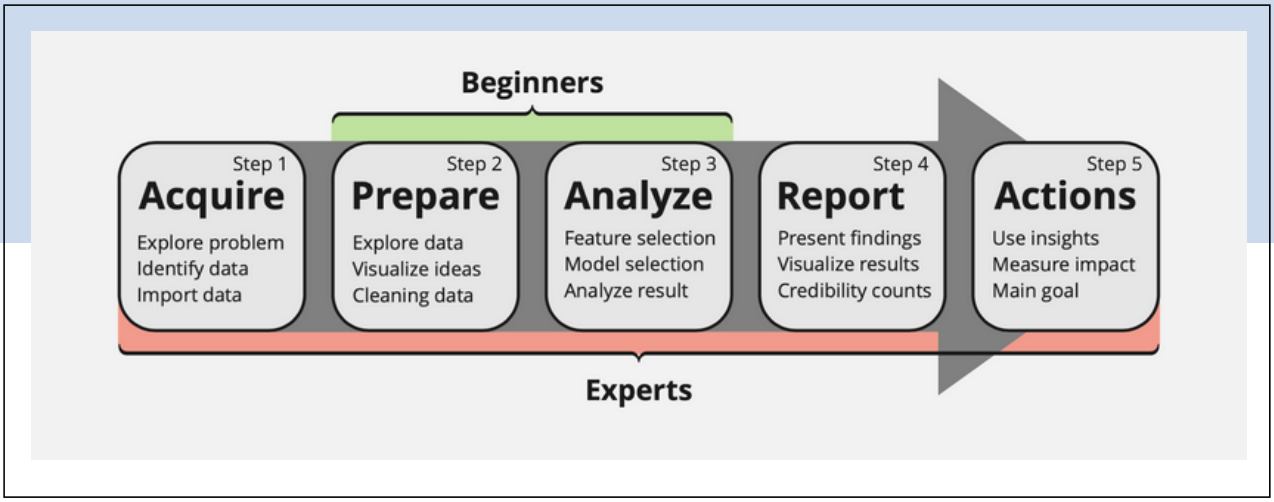
# Expert Data Science Blueprint

## Data Science Workflow

# What Makes This Data Science Course Different

- Expert Data Scientists focus on creating valuable actionable insights for clients.
- Beginner Data Scientists focus on covering the biggest tech stack.
- Experts know they need to understand the problem from start, to get the right data, and create client value.
- Beginners do not understand how each step in the Data Science Workflow is crucial to add value to the next step.




## Data Science Workflow

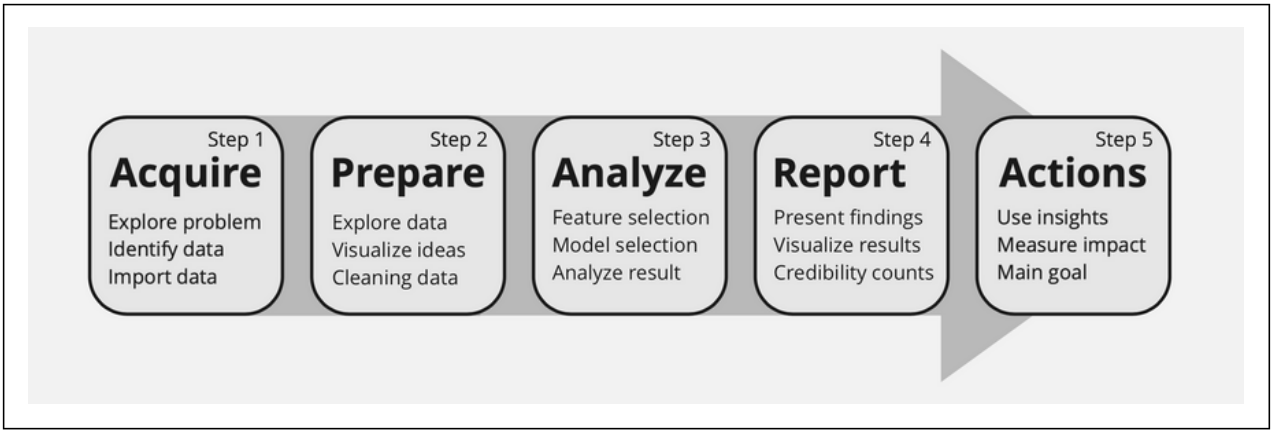
Mastering the Data Science Workflow is crucial, along with the right tools, to become an Expert Data Scientist. This course will cover all you need to start your journey toward Data Science Mastery.

At the end of the course, you will get a template covering all aspects to ensure your Data Science Project follows this flow and is done effectively with Python code using the right libraries.

# Data Science Course Curriculum

This is a 12 hours full Expert Data Science course. We focus on getting you started with Data Science with the most efficient tools and the right understanding of what adds value to a Data Science Project.

Most use too much time to cover too many technologies without adding value and end up creating poor-quality Data Science projects. You don't want to end up like that! Follow the Secret Data Science Blueprint, which will give a focused template covering all you need to create successful Data Science Projects.




- Data Science Workflow
- Data Visualization
- pandas for Data Science
- Data Sources: Web Scraping, Databases, CSV, Excel & parquet files
- Where to find data
- Join (combine) data
- Statistics you need to know
- Machine Learning Models
- Cleaning Data
- Feature Scaling
- Feature Selection
- Model Selection

## HOW TO GET STARTED

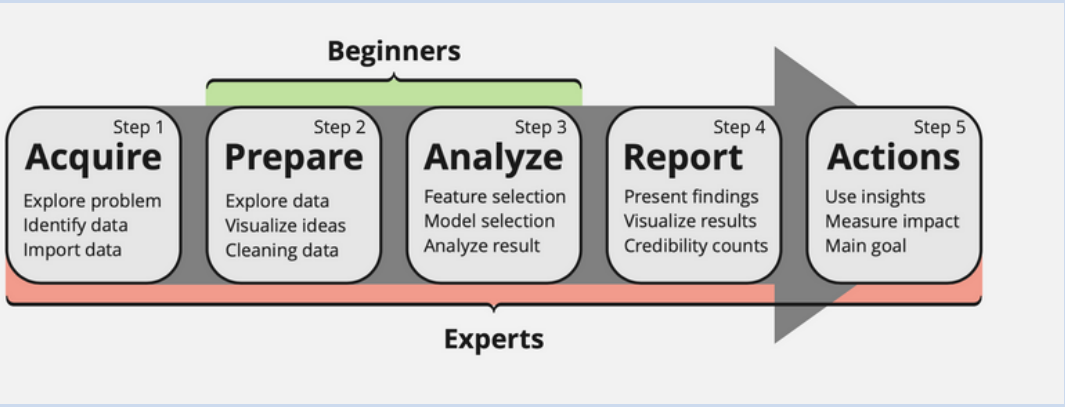
- Download all Jupyter Notebooks from the repo (**zip-file-download**).
- Unzip the download (main.zip) appropriate place.
- Launch Ananconda and start JuPyter Notebook (**Download** )
- Open the first Notebook from the download.
- Go to **LESSON 00 on** the next page

# LESSON 00


## Introduction to the Data Science Workflow

---

This will give you an understanding of what Data Science is and how to become successful. How to focus your effort to get the fastest results and not waste time on learning all possible technologies for Data Science.




### Lesson Objective

- What makes a successful Data Scientist?
- Why Data Science?
- How did Data Science start – this will surprise you!
- How the Data Science Workflow works.
- What skills does a Data Scientist need?
- Beginner vs Expert Data Scientist.
- Data Science Workflow used on real data with Python code.
- A project to try out yourself – make your first Data Science project.
- A solution to the project with code.

# LESSON 01


## Data Visualization for Data Science

---

Data Visualization for Data Science is not just how to present your data, it is about data quality, and exploring data to get an understanding of the nature of the data.
Data visualization helps you to understand data fast. Our human brain is not good at understanding rows of numeric data. But when we are visually presented with data we absorb information quickly. It improves our insights into data and enables us to make faster decisions.

.

**Lesson Objective**

- Understand the power of Data Visualization.
- How visualization enables us to see patterns in data.
- Learn how to use pandas and Matplotlib.
- What is Data Quality and how visualization helps you?
- How to spot wrong data entries.
- Identify outliners with visualization.
- How to Explore data with visualization.
- How to add a title, set labels, and adjust the axis.
- Use DataFrames and compare data.
- Create line charts, bar plots, histograms, pie charts, and scatter plots.
- Create annotation on the charts.

# LESSON 02


## pandas for Data Science

---

When working with tabular data (spreadsheets, databases, etc) pandas is the right tool with a built-in Data Structure.
pandas makes it easy to acquire, explore, clean, process, analyze, and visualize your data inside the DataFrame (pandas Data Structure).
pandas comes with a big framework of tools – which can be intimidating. In this lesson, we will break down what you need, how to find help, and how to work with pandas DataFrames.

**Lesson Objective**

- A brief introduction to pandas DataFrames
- What kind of data DataFrames are used for?
- Where to find further documentation
- The pandas' Cheat Sheet
- How to work with data in DataFrames
- Installing and importing pandas
- Reading CSV files with pandas into a DataFrame
- View data in a DataFrame
- Using index, columns, dtypes, shape, and length on a DataFrame
- Slicing a DataFrame on columns and rows.
- Creating new columns in a DataFrame
- Filtering and grouping data in a DataFrame
- Converting columns to DateTimes and changing string columns

# LESSON 03


## Web Scraping

---

Web Scraping is not only fun – it enables you to get data from any webpage and do your own analysis of the data.
With the pandas' library in Python, you can do that in two steps and have the data prepared for further processing.
First – get the URL of the page and use pandas read_html to parse the data from the webpage into a list of DataFrames.
Second – Data Wrangling – which means transforming the data to be in the right format for further processing. That can be extracting numeric values from entries like: “$ 1,234,567” or converting dates to date objects.

**Lesson Objective**

- Understand what Web Scraping is
- How to use the pandas' library to Web Scrape
- Understand legal issues with Web Scraping
- Demonstrate the challenge with data after Web Scraping
- Learn what Data Wrangling is
- How to see the data types of Web Data
- How to convert data to numeric values in a DataFrame
- To extract values from strings from DataFrame
- Error handling while converting DataFrames

# LESSON 04


## Master Databases with pandas

---

Learn how to get data from Databases and into a pandas DataFrame. We will also learn how to join data from tables into single DataFrames.
We will cover what Relational Databases are, and how it models data in rows in columns in a series of tables. You will learn how databases resemble a collection of DataFrames or Excel sheets of data.
You will get the most common SQL statements (Structured Query Language) needed to get data from the database to DataFrame. It is actually less than you think.
We will work on real SQLite databases, but you will know how to connect to other databases.

**Lesson Objective**

- What is a database?
- Understand the Relational database model.
- Most used SQL queries.
- How to use a database connector.
- Sqlite3 connector to connect to SQLite database.
- How to use sqlite3 with SQLite files.
- List all tables in the database.
- How to use database connector with pandas DataFrames.
- SQL join syntax to join multiple tables into one DataFrame.
- Use Folium to plot data on a map.

# LESSON 05


## Read CSV, Excel and Parquet files

---

You will learn how to read data from CSV, Excel, and Parquet files into pandas DataFrames.
This includes what the file formats contain, how they differ, and what format you should store your data in – which is context-dependent.
Also, you will learn the most commonly used arguments for each method to read data into pandas' DataFrame.
Finally, you will get a collection of great places to find data online.
In the project, we will be given a Data Science problem and we need to find a great place to find data to answer the question. We will learn a great lesson in how different representations of the data can give different views.

**Lesson Objective**

- What a CSV file is.
- How to read a CSV file.
- The most common arguments to the read_csv method.
- How to read an Excel file.
- Learn how to set the index column when reading Excel files.
- What a Parquet file is.
- Why use Parquet files over other types?
- How to read a Parquet file.
- Where to find data online.
- How different presentations can tell different stories.

# LESSON 06


## Combine DataFrames

---

You will learn how to combine data from two pandas DataFrames into one using Merge, Join, and Concat.
This will teach you how to add additional data to an existing dataset. As an example, you will try to combine metadata to the total population dataset by country from the World Bank.
The dataset comes with additional metadata, which you will merge into the main dataset. This adds metadata like “region” and “income group” to each country.
In the project, you will explore if GDP is correlated to SPI (Statistical Performance Indicators). In general, you will look into what the SPI tells us, and if there are regional differences in the SPI score, before looking at the correlation.

### Lesson Objective

- How to combine 2 DataFrames into 1.
- See 3 different ways: Merge, Join, and Concat.
- Explore Merge with real data from the World Bank.
- Understand how adding metadata can enrich data analysis.
- Group data by new metadata.
- Understand SPI.
- See if GDP correlates with SPI.
- Is SPI different in regions?

# LESSON 07


## Statistics for Data Science

---

In this lesson, you will learn the most common statistics you need for Data Science.
Statistics is one of the areas where most get lost and scared because there is so much to learn and it is difficult to know what is relevant.
Here we will show you what you need to understand – and surprisingly, it is not that difficult to understand. At the end of this video, you will know what is the most important statistics, what mean is, how to use mean and groupby in DataFrames, know what the standard deviation is and how to use it, what insights describe a DataFrame gives, how you read box plots and some insights into correlations.

**Lesson Objective**

- What are the most important statistics?
- Learn what mean is.
- How to use mean with groupby on DataFrames.
- Understand standard deviation (std).
- What does standard deviation (std) and mean tell you about data?
- How to use describe on DataFrames and what it tells you.
- Interpret box plots.
- Use box plots on DataFrames with groups.
- Understand the correlation of data.
- How to use statistics to optimize your salary.

# LESSON 08


## Get Started with Linear Regression

---

Do you like soccer? Do you want to learn how to predict Soccer Player Ratings? To do that we need to learn about Linear Regression, which is a Machine Learning model used to predict continuous values. Linear Regression describes the relationship between variables.
We will use Sklearn (scikit-learn) for the Linear Regression and demonstrate how it should be understood and visualize how it works. Also, we will look into how to measure the quality of the Linear Regression model.
The project will explore the European Soccer Database, which is a very popular dataset on Kaggle. We will make a Linear Regression model to predict Player Ratings, based on the other features.

**Lesson Objective**

- What is Linear Regression?
- Similarities between Linear Regression and Correlation.
- Visually understand what Linear Regression can solve.
- Show how and what Linear Regression does.
- Use Sklearn LinearRegression model.
- Visualize the prediction.
- Understand what R-squared tells us.
- Work with European Soccer Database.
- Predict Player Ratings with Linear Regression.
- Inspire to a bigger project.

# LESSON 09


## Missing data and Data Cleaning

---

What to do with NaN? Just use dropna?
In this lesson, you will learn the impact of just deleting rows with missing data and how to deal with it. You will learn to replace missing data with mean values and how you can use interpolation on time series data.
This is part of cleaning data and starts with an understanding of data quality. Common issues with data quality can be data outliers and duplicates. We will also explore and show how to deal with that.
In the project, we will look at how we can use interpolation on a weather dataset to restore missing values. This will show you how big an impact it can have on the accuracy of your model.

**Lesson Objective**

- Understand data quality.
- How to improve data quality.
- How to deal with missing data.
- This includes replacing and interpolating data.
- How to find outliers with visualization.
- Identifying data in wrong units.
- Dealing with duplicates.
- Demonstrating the impact of dealing with missing data.
- Real project with data interpolation.

# LESSON 10


## Machine Learning vs Classical Computing

---

Why is Machine Learning so brilliant?
How would like to learn something new? Would you like to get a list of 100 specific instructions to follow specifically? Or would you prefer to figure it out yourself to get the desired outcome?
The first approach is actually classical computing, while the second is the Machine Learning approach. It seems that computers are better at figuring out how to solve some problems than we humans are at describing how to do it
In this lesson you will learn how Machine Learning works, we will specifically dive into Supervised Learning and Classify the Iris Flower Dataset.

**Lesson Objective**

- How does Classical Computing work?
- What makes Machine Learning great at certain tasks?
- What is the different Machine Learning approaches?
- Understand Supervised, Unsupervised, and Reinforcement Learning.
- Learn the Machine Learning work process.
- The difference between learning and prediction phases.
- Dive into Supervised Learning.
- Demonstrate how it works on Iris Flower Dataset.
- Understand the training and test dataset division.
- Find the most important features.
- Visualize feature impact and classification
- Model accuracy

# LESSON 11


## Feature Scaling

---

What is Feature Scaling? And how can it help you?
In this lesson, you will learn about the two types of Feature Scaling: Normalization and Standardization. You will see the difference visually to understand the different approaches. Also, you will learn when you need to feature scale.
Feature scaling is also a great idea when you want to compare results.
You will see box plots of the different approaches and see the impact on real-life weather data. The data will demonstrate the problem easily and visually, and finally the accuracy of the models, you will create in this lesson.

**Lesson Objective**

- What is Feature Scaling?
- How Feature Scaling transforms data into similar ranges.
- Understand the difference between Normalization and Standardization.
- The mathematical description of the approaches.
- That distance-based algorithms are the most sensitive.
- KNN, K-means, and SVM need Feature Scaling.
- Use sklearn MinMaxScaler for Normalization.
- Use sklearn StandardScaler for Standardization.
- Visualize the different approaches with box plots.
- Compare the accuracy of Feature Scaling approaches.

# LESSON 12


## Feature Selection

---

Feature Selection seems like an advanced topic most beginners skip while training Machine Learning Models.
This is a wrong approach. First of all, Feature Selection is not difficult when you know how to do it. Second of all, this will give you higher accuracy, simpler models, and reduce the risk of overfitting.
There is a long list of Feature Selection Techniques. This includes simple Filter methods, Wrappermethods, and Embedded methods.
Luckily, you get the biggest impact by using simple approaches. In this lesson, we will explore Filter methods like removing constant and quasi–constant features, and removing correlated features. With Wrapper methods, you will learn Forward Selection.

### Lesson Objective

- What is Feature Selection?
- How Feature Selection gives you higher accuracy.
- Feature Selection gives simpler models.
- It minimized the risk of overfitting the models.
- Learn the main Feature Selection Techniques.
- Filter Methods are independent of the model.
- This includes removing Quasi-constant features.
- How removing correlated features improves the model.
- Wrapper Methods are similar to a search problem.
- Forward Selection works for Classification and Regression.

# LESSON 13


## Model Selection

---

How do you know which model to use for your Machine Learning project? If you choose the wrong one, will it make you look inexperienced?
Most want to find out what is the BEST model for the project. We need to understand that all models have predictive errors.
What we should seek is a model that is good enough. On a top level, the problem type we want to solve will guide us from which category of models we should choose. The next level will be using a model selection technique to find the right one from the subset of models.
In this lesson, you will learn about probabilistic measures and resampling methods for selecting models.

**Lesson Objective**

- There is no best ML model for a problem.
- All ML models have predictive errors.
- How the problem type narrows down on ML models.
- Using the Sklearn cheat sheet.
- What are Model Selection Techniques?
- Probabilistic Measures score performance and complexity of the model.
- That Resampling Methods are split into sub-train/test and score it.
- How to convert to Categories.
- Difference between cut() and qcut().
- How to calculate the Accuracy Score.

# LESSON 14


## The Ultimate Data Science Workflow Template

---

When it comes to creating a good Data Science Project you will need to ensure you cover a lot of aspects. This template will show you what to cover and where to find more information on a specific topic.
The common pitfall for most junior Data Scientists is to focus on the technical part of the Data Science Workflow. To add real value to the clients you need to focus on more steps, which are often neglected.
This guide will walk you through all steps and elaborate and link to in-depth content if you need more explanations.
In the **project,** you will try the template and get demonstrated how to use a great example from **IMDB**.

### Lesson Objective

- Master the Data Science Workflow.
- Understand what is important to create valuable Data Science projects.
- Why understanding the problem is a crucial first step.
- Great sources to find additional data.
- Where to find in-depth knowledge from each step.
- How to explore data to understand it better.
- What statistics can help you with this?
- Choosing a model for the project.
- How to train and test a model.
- Making great presentations of your findings.
- Create insights and measure impact.

# Congratulations

**You have created 15 projects with Python for Data Science!**

You learned a lot

- Data Science Workflow
- Customer value
- DataFrames
- Visualization
- Data Quality

...and much more.

**It is the first step**

**You can actually create valuable Data Science projects now!**

- practice makes you better
- specialize in what you find interesting

I started programming at 12, got a Ph.D. in computer science, coded professionally for almost 15 years, and I still need help and to learn new stuff all the time.

One step at a time... but enjoy the learning journey!

*Rune*