

- **Data Analyst Capstone Project**
- **Week 1 : Data Collection**
- **Welcome to the course**
- **Course Introduction**
- **3 min**

- Congratulations on making it this far! Successfully completing this Capstone project will earn you the Data Analyst Professional Certificate. This Capstone provides you a practical hands on experience to demonstrate all of the skills you have picked up so far in this Professional Certificate program
- **Introduction to Capstone Project**

Introduction To Capstone Project

You have recently been hired as a Data Analyst by a global IT and business consulting services firm that is known for their expertise in IT solutions and their team of highly experienced IT consultants. In order to keep pace with changing technologies and remain competitive, your

organization regularly analyzes data to help identify future skill requirements.

As a Data Analyst, you will be assisting with this initiative and have been tasked with collecting data from various sources and identifying trends for this year's report on emerging skills.

Your first task is to collect the top programming skills that are most in demand from various sources including:

- Job postings
- Training portals
- Surveys

Once you have collected enough data, you will begin analyzing the data and identify insights and trends that may include the following:

- What are the top programming languages in demand?
- What are the top database skills in demand?
- What are the popular IDEs?

You will begin by scraping internet web sites and accessing APIs to collect data in various formats like .csv files, excel sheets, and databases.

Once this is completed, you will make that data ready for analysis using data wrangling techniques.

Syllabus

Module 1: Data Collection

- Collecting Data Using APIs
-
- Collecting Data Using Webscraping
-
- Exploring Data

Module 2: Data Wrangling

- Finding Missing Values
-
- Determine Missing Values
-
- Finding Duplicates
-
- Removing Duplicates
-
- Normalizing Data

Module 3: Exploratory Data Analysis

- Distribution

-
- Outliers
-

- Correlation

Module 4: Data Visualization

- Visualizing Distribution of Data
-
- Relationship
-
- Composition
-
- Comparison

Module 5: Dashboard

Creation

- Dashboards

Module 6: Presentation of Findings

- Final Presentation
- **Collecting Data Using APIs**
- **(Optional) Hands-on Lab 0: Getting Started with Watson Studio**
- **<https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DA0321EN-SkillsNetwork/labs/module%201/Lab%20-%20Getting%20Started%20with%20Watson%20Studio/instructions.md.html?origin=www.coursera.org>**
-

- **Claim IBM Cloud Feature Code**
- **Unavailable**
- **(Optional) Hands-on Lab 1: Review Of Accessing APIs**

In the Python for Data Science course you learned to use REST APIs.

API access details

Locked

Hands-on Lab 2: Collecting Data Using APIs

We are using two notebooks in this lab. First is Jobs_API which you just need to execute (No changes to be made here).

Second is

Collecting_Jobs_data_Using_API notebook in which you need to write the code and find the required results. Right click and open [this link](#) to download the Jupyter notebook. Then open it using Watson Studio and follow the instructions in the notebook to complete the lab.

Note: Both the notebooks are related to each other, so running Jobs_API notebook before the

Collecting_Jobs_data_Using_API notebook is must. Also, remember to upload both the notebooks in the same project in Watson.

Graded Quiz: Collecting Data

Which python module helps you to easily access an API?

Requests

How many job postings exist for Python? 51

Which programming language has the maximum number of job postings?

C

Collecting Data Using Webscraping

Hands-on Lab 3: Review Of Web Scraping

<https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DA0321EN-SkillsNetwork/labs/module%201/Web%20Scraping/Web-Scraping-Review-Lab.ipynb>

Hands-on Lab 4: Collecting Data Using Web Scraping

<https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DA0321EN-SkillsNetwork/labs/module%201/Web%20Scraping/Web-Scraping-Lab.ipynb>

Graded Quiz: Web Scraping

Which language are developers paid the most according to the output of the web scraping lab?

Swift

Which of the following code segments will help to scrape all images from a html page?

```
soup.find_all("img")
```

Exploring Data Set

Hands-on Lab 5: Explore the Data Set

Right click and copy [this link](https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DA0321EN-SkillsNetwork/labs/module%201/Exploring%20the%20dataset/M1ExploreDataSet-lab.ipynb) to the Jupyter notebook for this lab. Then open it using Watson Studio and follow the instructions in the notebook to complete the lab.

<https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DA0321EN-SkillsNetwork/labs/module%201/Exploring%20the%20dataset/M1ExploreDataSet-lab.ipynb>

Graded Quiz: Exploring Data

How many rows are in the data set

11552

How many columns are in the data set?

85

What type of column is "Age"?

Numerical

Week 2

Data Wrangling

Assignment overview

Data Wrangling or Munging is a process in which we clean up the data set and make it ready for data analysis. In this assignment you will perform the following tasks:

1. Identify duplicate rows in the data frame.
2. Remove duplicate rows from the dataframe.
3. Find the number of missing values for all columns.
4. Find the value counts for the column "Employment".
5. Normalize the data using two existing columns.

Jupyter Notebook For Labs In This Module

Right click and copy [this link](#) for the Jupyter notebook to be used for all of the labs in this module.

<https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DA0321EN-SkillsNetwork/labs/module%202/M2DataWrangling-lab.ipynb>

Graded Quiz Duplicates

How many duplicate rows are there in the dataset?

154

How many duplicate values are there in the column Respondent

154

Graded Quiz Removing Duplicates

After removing the duplicate rows, how many rows are there in the dataset?

11398

After removing the duplicate rows, how many unique rows are there in the column Respondent?

11398

Graded Quiz missing values

After removing the duplicate rows, how many blank rows are there under the column EdLevel?

112

After removing the duplicate rows, how many rows are missing under the column Country?

0

Graded Quiz Imputing Missed values

What is the majority category under the column Employment?

Employed full-time

Under the column "UndergradMajor", which category has the minimum number of rows?

Health Science

The column 'ConvertedComp' contains the annual compensation of the survey respondents. What is the best approach to impute the missing values in this column?

Median

Graded Quiz Normalising Data

How many unique values are there in the CompFreq column?

3

After removing the duplicate rows, how many respondents are being paid yearly?

6073

What is the median NormalizedAnnualCompensation?

100000

Week 3 : Exploratory Data Analysis

Assignment Overview

After cleaning the dataset, your next step is the analysis. In this stage you will become more familiar with the data set and it will start to take shape. In this assignment you will:

1. Plot a distribution curve, and histogram.
2. Find the median, and outliers of particular columns.
3. Compute the Inter Quartile Range.
4. Find out the upper and lower bounds, and find correlations between numerical columns.
5. Create a new dataframe.

Jupyter Notebook For Labs In This Module

Right click and copy [this link](https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DA0321EN-SkillsNetwork/labs/module%203/M3ExploratoryDataAnalysis-lab.ipynb) for the Jupyter notebook to be used for all of the labs in this module.

<https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DA0321EN-SkillsNetwork/labs/module%203/M3ExploratoryDataAnalysis-lab.ipynb>

Graded Quiz Distribution

What is the median Age of survey respondents? 29

What is the median ConvertedComp of respondents who have identified themselves as 'Woman'? 57708

Based on the histogram, which age group has the highest number of survey respondents?

24-32

Graded Quiz Handling Outliers

What is the median ConvertedComp before removing outliers?57745

What is the median ConvertedComp after removing outliers?52704

Based on the boxplot of 'Age' how many outliers do you see below Q1?

0/1/2

What is the mean ConvertedComp after removing outliers?

59883

Graded Quiz Correlation

Which column has a negative correlation with "Age"?

CodeRevHrs

column has the highest correlation with "Age"?

ConvertedComp

Week 4 : Data Visualization

In this assignment you will now show off your visualization skills by working with the Stack Overflow Developer Survey 2019 dataset. In this assignment you will:

1. Create a histogram to show the distribution of data.
2. Create different plots such as a scatter, bubble or boxplot.
3. Create a pie chart, bar chart, and stacked chart to show medians and counts.

Jupyter Notebook For Labs In This Module

Right click and copy [this link](https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DA0321EN-SkillsNetwork/labs/module%204/M4DataVisualization-lab.ipynb) for the Jupyter notebook to be used for all of the labs in this module.

<https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DA0321EN-SkillsNetwork/labs/module%204/M4DataVisualization-lab.ipynb>

Graded Quiz Visualizing Distribution Of Data

Based on the boxplot of the column "Age", where do you see the outliers?

Above Q3

Graded Quiz Visualizing Relationship

Based on the scatter plot of Age and WorkWeekHrs , what kind of correlation do you see between them?

Difficult to clearly identify from the scatter plot.

Graded Quiz Visualizing Composition of Data

In the pie chart of top 5 databases, respondents wish to learn next year, what is the percentage of MongoDB?

20.9

In the list of most popular languages respondents wish to learn next year, what is the rank of Python?

3

How many respondents indicated that they currently work with 'SQL'?

7106

How many respondents indicated that they work on 'MySQL' only?

657

876

474

Graded Quiz Visualizing Comparison of Data

Majority of the survey responders are

Developer - full-stack

Based on the Line Chart of Age and median ConvertedComp for the age group 25-30, what kind of correlation do you see between Age and median ConvertedComp?

Positive

Week 5 : Building a Dashboard

Assignment Overview

In this assignment you will continue working with the Stack Overflow Developer Survey 2019 data to create a dashboard. In this dashboard you will create the following:

A Current Technology Usage tab containing:

1. Top 10 Languages
- 2.

3. Top 10 Databases
- 4.
5. Platforms
- 6.
7. Top 10 WebFrames

A Future Technology Trends tab containing:

- Top 10 Languages desired for the next year
-
- Top 10 Databases desired for the next year
-
- Desired platforms for the next year
-
- Top 10 WebFrames desired for the next year
-

A Demographics tab containing:

- Respondent classified by gender
-
- Respondent count for countries
-
- Respondent count by age
-
- Respondent count by gender and classified by education level

Hands-on Lab 22: Getting Started
With Cognos Dashboard Embedded

<https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DA0321EN-SkillsNetwork/labs/module%205/Lab%20-%20Getting%20Started%20with%20Cognos%20Dashboard%20Embedded/instructions.md.html?origin=www.coursera.org>

Building A Dashboard With Cognos Dashboard Embedded

<https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DA0321EN-SkillsNetwork/labs/module%205/Peer%20Graded%20Assignment%20-%20%20Building%20a%20dashboard%20with%20Cognos%20Dashboard%20Embedded/instructions.md.html?origin=www.coursera.org>

Peer-Graded Assignment: Dashboards

Dashboard Link
<https://dataplatform.cloud.ibm.com/dashboards/e36a0061-f236-4a37-b128-ce8c060a1dbd/view/5529a17964ea3df16fe2eee4079b7f037561200ee6bb8b0383817b495a327297a9611095c87d4c5988450135f0e81151c1>

Peer-Graded Assignment:
Dashboards PDF

Module 5

Assessment Information

Unlike other modules there is no quiz in this module. However as part of the peer graded assignment in the next module you will be sharing a permanent read-only link for your Cognos dashboard and your peers will grade you according to the assessment criteria. The assessment for this evaluation accounts for 10% of your overall grade for this course. In addition you will also need to include screen shots of your dashboard tabs in the final presentation

Week 6

How To Present Your Findings

Elements Of A Successful Data
Findings Report
Video

While finding and cleaning data is an important first step in data analysis, a concept can be lost if you are not able to organize and represent the findings effectively to your audience. In this video, you will learn how to represent your findings by focusing on specific elements to create a successful data findings report.

In this video, we learned about the important elements in creating a successful data findings report. In the next video, we will learn the best practices when presenting your findings.

Structure Of A Report

<https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0101EN-SkillsNetwork/labs/Module%203/ReadingReportStructure.md.html?origin=www.coursera.org>

Best Practices For Presenting Your Findings

Video

Okay, you've spent weeks, maybe months, studying the data and the time has come to report your findings. The questions have been answered,

and you feel good about the story.

In this video, we learned about creating a data-driven presentation that will keep the audience engaged and how to deliver a clear and concise message.

(Optional) Hands-on Lab 24: Getting Started With PowerPoint For The Web

<https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DA0321EN-SkillsNetwork/labs/module%206/Lab%20-%20Getting%20Started%20with%20PowerPoint%20for%20the%20Web/instructions.md.html?origin=www.coursera.org>

(Optional) Hands-on Lab 25: Basics of PowerPoint

<https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DA0321EN-SkillsNetwork/labs/module%206/Lab%20-%20Basics%20of%20PowerPoint/instructions.md.html?origin=www.coursera.org>

(Optional) Hands-on Lab 26: Save your PowerPoint Presentation as PDF

<https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DA0321EN-SkillsNetwork/labs/module%206/Lab%20->

<https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DA0321EN-SkillsNetwork/labs/module%206/Lab%20-%20Getting%20Started%20with%20Save%20your%20PowerPoint%20Presentation%20as%20PDF/instructions.md.html?origin=www.coursera.org>

<https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DA0321EN-SkillsNetwork/labs/module%206/Lab%20-%20Getting%20Started%20with%20PowerPoint%20for%20the%20Web/capstone-story-template.pptx>

Assignment Overview

For your final assignment you will create a presentation based on the analysis of your data conducted in previous modules. Your presentation will develop into a story of your analysis that should be compelling and easy to understand. A PowerPoint template has been provided to get you started however you are free to add additional slides, charts and tables.

Exercise: Preparing Your Presentation

Download the Microsoft PowerPoint [presentation template](#) from the link below:

<https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DA0321EN-SkillsNetwork/labs/module%206/Lab%20-%20Getting%20Started%20with>

Note: If you do not have a desktop version then follow the instructions from the previous lab to get access to PowerPoint online.

1. Complete the slides with relevant details (You are free to add additional slides, charts, tables).
2. Add any additional unused charts or tables to the appendix that you think are relevant to your analysis.
3. Convert your final presentation into .pdf format.
4. As part of your grade for the final assignment, you will be asked to upload your presentation for peer review.
5. You will also need to review the final presentation for at least one other peer and grade them before your final grade will be available.

Peer-graded Assignment

6. PPT
7. <https://coursera-assessments.s3.amazonaws.com/assessments/1639503460859/b6fb47d4-702a-4572-96d5-5bd0811364fd/SWAROOP%20capstone.pdf>

**Peer Graded Assignment
Questions PDF**