**Module 16 – Big Data**

**In this module we will explore the big data ecosystem including Hadoop, the four Vs of Big Data, MapReduce, and Spark. We will also cover natural language processing. This is when we convert the language we speak into something computers can understand. Natural language processing is a critical big data concept and skill set because there’s so much spoken and written language that needs to be translated. We use natural language processing in spam filters, spell check, and voice recognition technology.**

**For this project I am the first data expert at BigMarket, a startup that helps businesses optimize their marketing efforts. And you’ve been hired just in time too. One of BigMarket’s clients has requested some pretty hefty analytics. This client $ellby is about to release a large catalog of products on a leading retail website. They want to know how the reviews of their products compare to the reviews of similar products sold by their competitors. They’re also interested in enrolling in a program that gives out free products to select reviewers but they want to know if it’s worth the cost. There thousands of reviews and they’re in words not numbers so you’ll need to translate them in order to analyze them. $ellby hired BigMarket because they believe in the power of big data.**

**For** this project you'll be partnering with Jennifer, an account manager at BigMarket. SellBy, your client, loves to talk about the power of big data, but Jennifer isn't a data expert. So you start off the project by giving her a quick overview of what big data actually is.

A good rule of thumb to apply is this: Data is considered big data when it exceeds the capacity of operational databases.

**Four Vs of Big Data**

There are four characteristics of big data:

* **Volume** refers to the size of data (e.g., terabytes of product information). For instance, a year's worth of stock market transactions is a large amount of data.
* **Velocity**pertains to how quickly data comes in (customers across the world purchasing every second). As an example, McDonald's restaurants are worldwide with customers buying food at a constant rate, so the data comes in fast.
* **Variety** relates to different forms of data (e.g., user account information, product details, etc.). Consider the breadth of Netflix user information, videos, photos for thumbnails, and so forth.
* **Veracity** concerns the uncertainty of data (e.g., reviews might not be real and could come from bots). As an example, Netflix would want to verify whether users are actively watching the shows, falling asleep, or just playing them in the background.

The four Vs of big data will help you determine when to migrate from regular data to big data solutions.

# **Big Data Technologies**

Apache Hadoop (Hadoop) is one of the most popular open source frameworks, with numerous technologies for big data. Google developed Hadoop to process large amounts of data by splitting data across a distributed file system.

We'll start with the three main components of Hadoop:

* **Hadoop Distributed File System (HDFS)** is a file system used to **store data** across server clusters (groups of computers). It is scalable (which means it handles influxes of data), fault-tolerant (handles hardware failure), and distributed (spread across multiple servers connected by a common core).
* **MapReduce** is a programming model and processing technique for big data. MapReduce enables processing the large amount of data spread across the cluster in the HDFS by performing the same task for each file system.
* **Yet Another Resource Negotiator (YARN)** manages and allocates resources across the clusters and assigns tasks.

Hadoop distributes for the storage and processing of data through a cluster, which is a group of connected computers that work together to store and perform tasks on a dataset.

Hadoop is quite difficult to set up. You need to set up all three main components across multiple machines, as well as make sure each one has sufficient resources and is configured for optimal performance. Because of this, it may not be the right technology for your startup. However, you know your client will ask about it—"Hadoop" is a popular buzzword, after all—so it's important to have a baseline knowledge of it.

**“/Users/rfnichol/OneDrive - COOPER TIRE & RUBBER COMPANY/Personal/Data Analytics Boot Camp/Module\_16\_Big\_Data”**

**Notes 16.1**

* The activities in this class will complement Lessons **16.1.1: What is Big Data?** through **16.4.3: Spark Functions**. The students will benefit from these activities if they‘ve progressed through these lessons, which cover the following concepts, techniques, and tasks:
* Big data overview
* Spark architecture
* Google Colab Notebooks
* Spark DataFrames and Datasets
* Spark transformations and actions

**Notes 16.2**

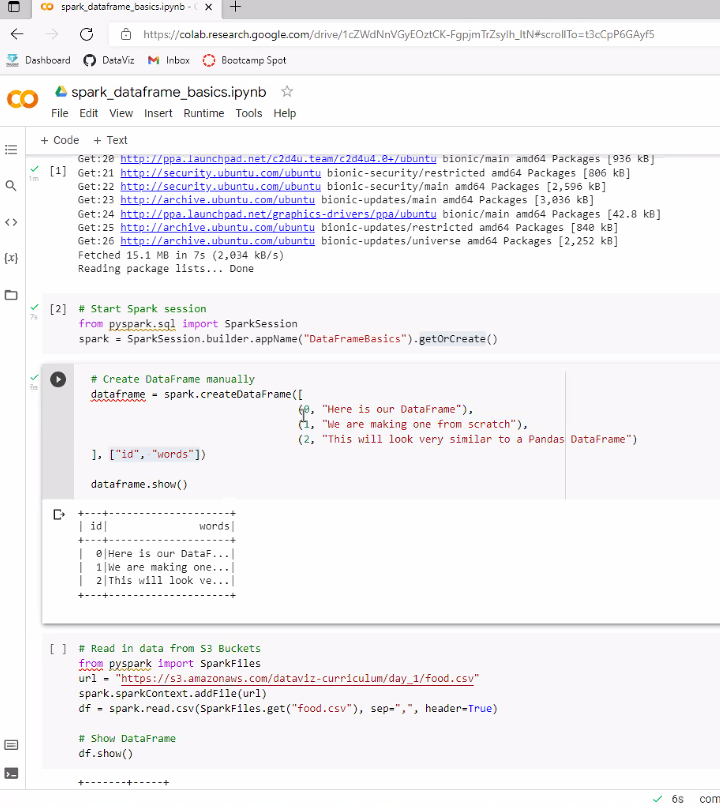
* The activities in this class will complement Lessons **16.7.1: Evaluate Amazon Web Services** through **16.9.3: Check AWS billing**. The students will benefit from these activities if they‘ve progressed through these lessons, which cover the following concepts, techniques, and tasks:
* Using AWS
* Creating a Relational Database in AWS
* Connecting with pgAdmin to an RDS instance
* Storing data on AWS S3
* Retrieving data from S3 and performing ETL to load into an RDS instance.
* Managing AWS billing.

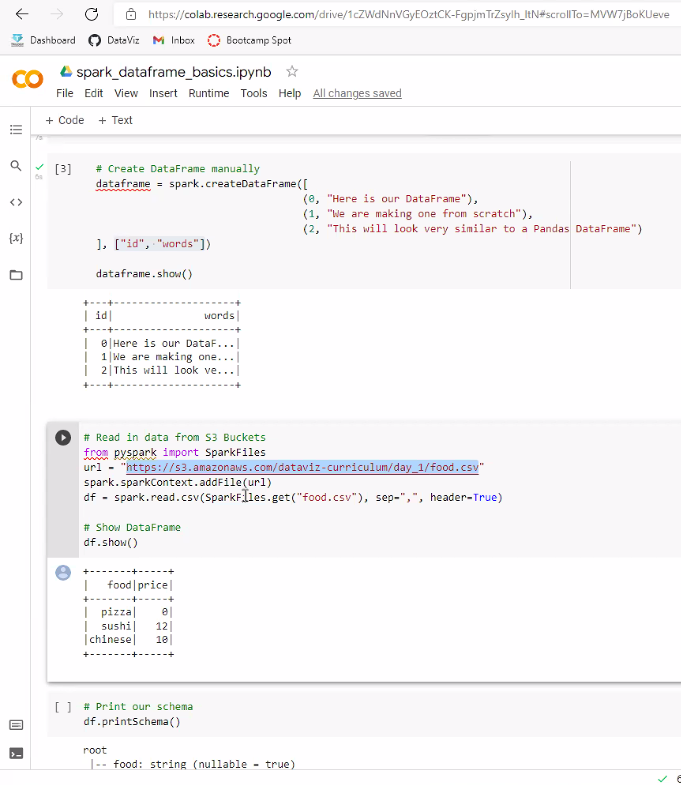
Challenge 16

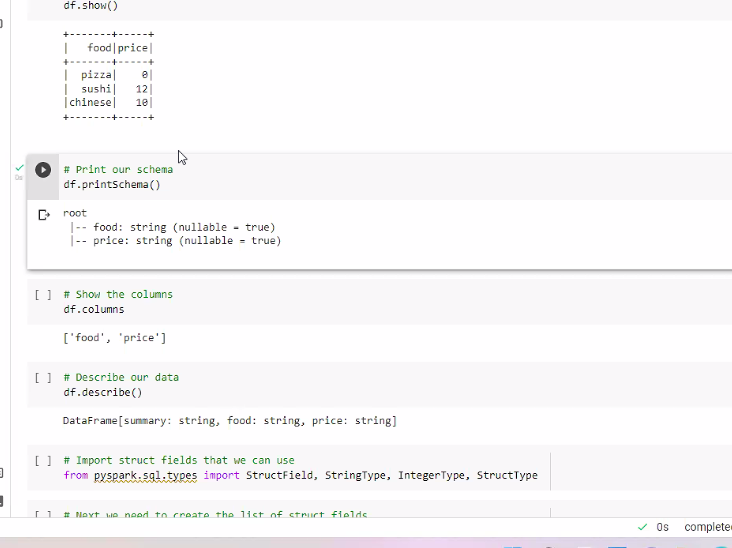
Reference

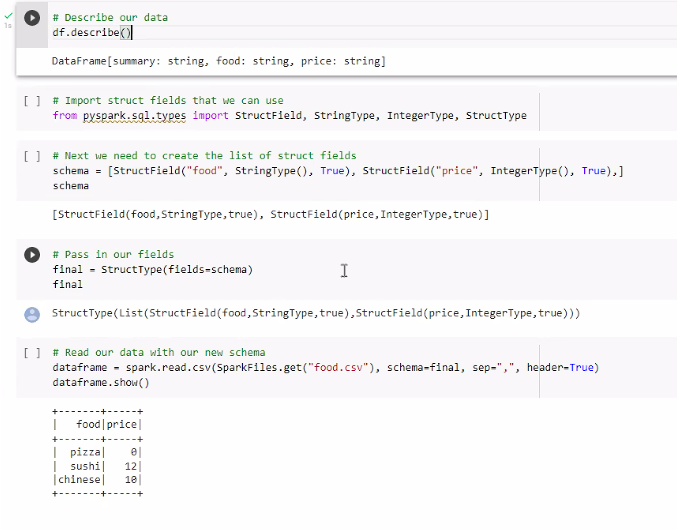
<https://github.com/Deving789/Amazon_Vine_Analysis/blob/main/Amazon_Reviews_ETL.ipynb>

<https://github.com/sahossain99/Amazon_Vine_Analysis>

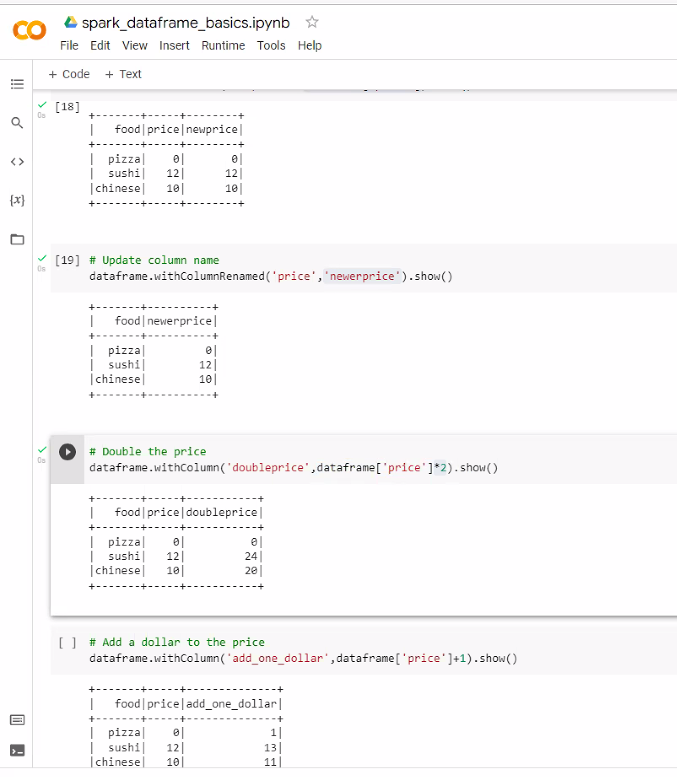


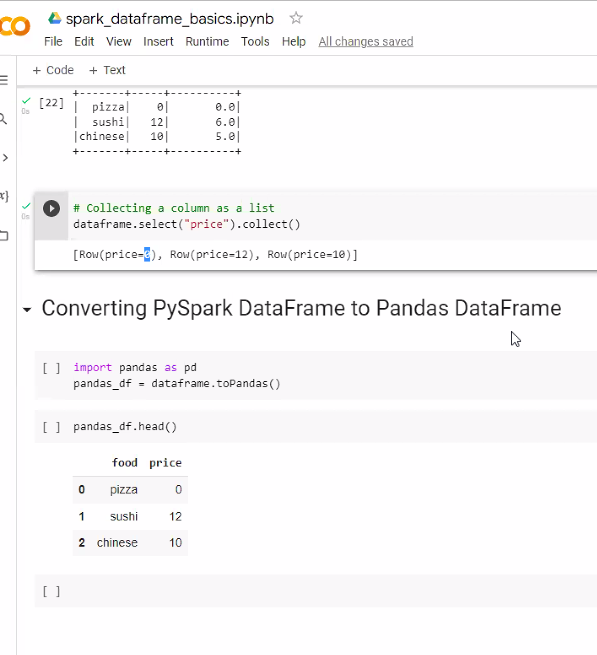


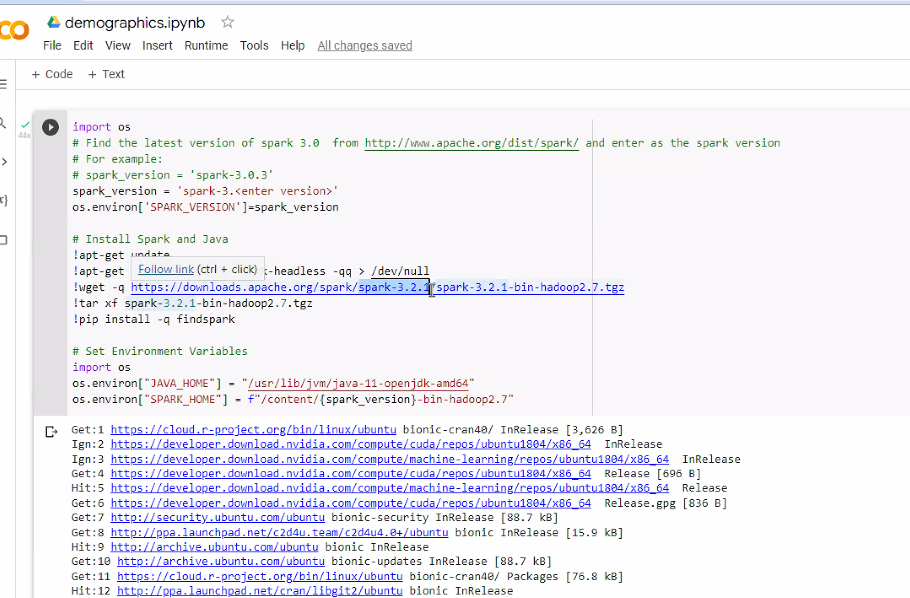


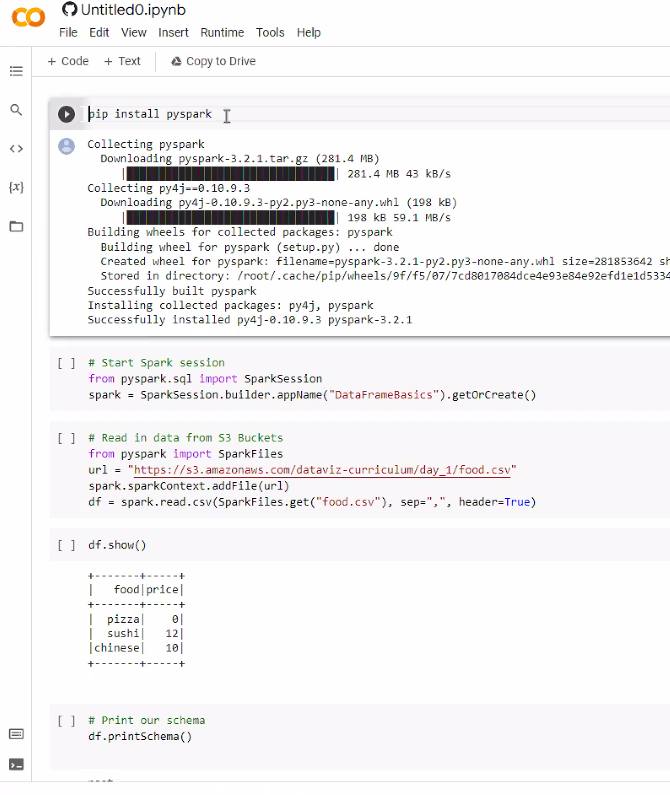


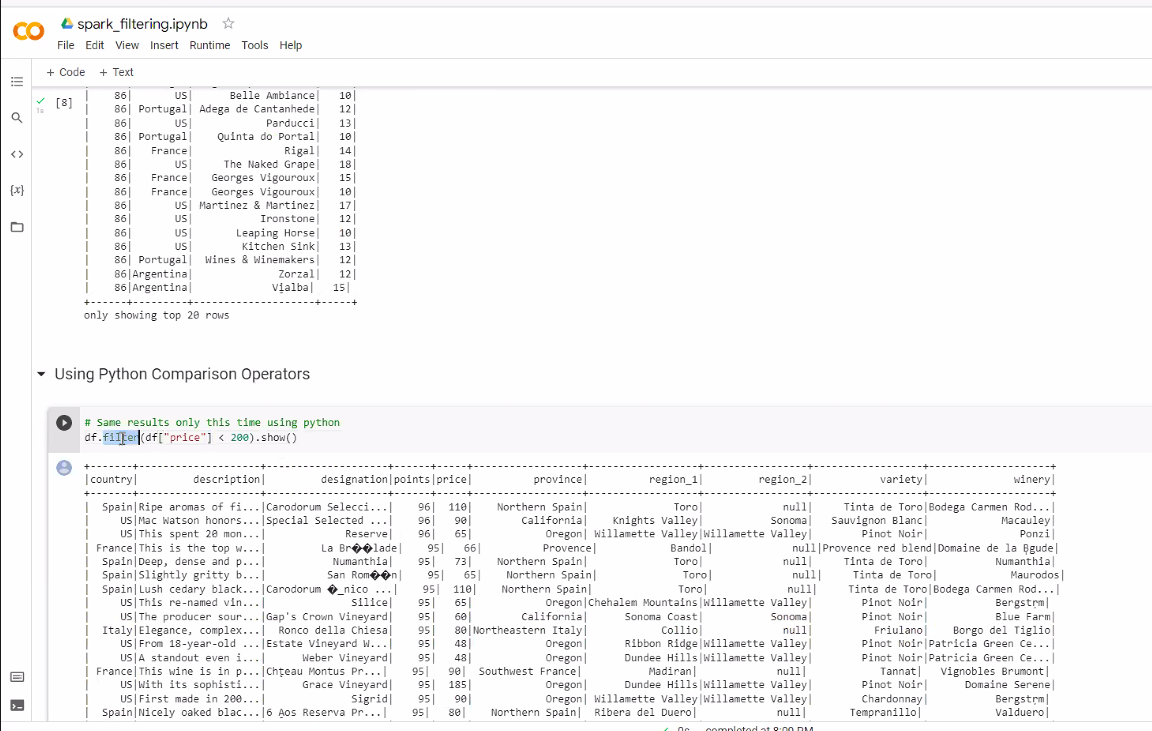
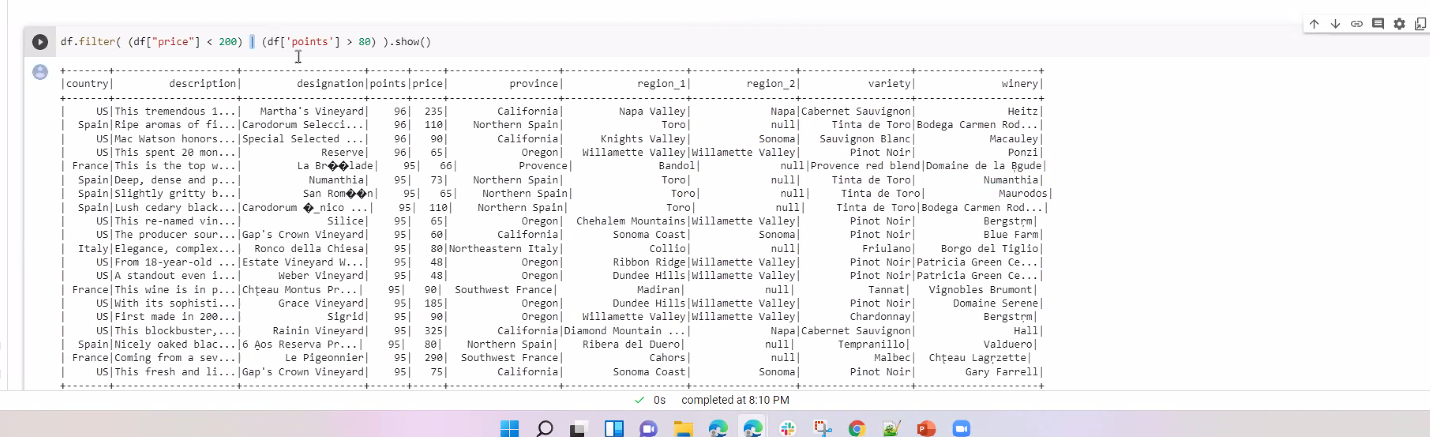


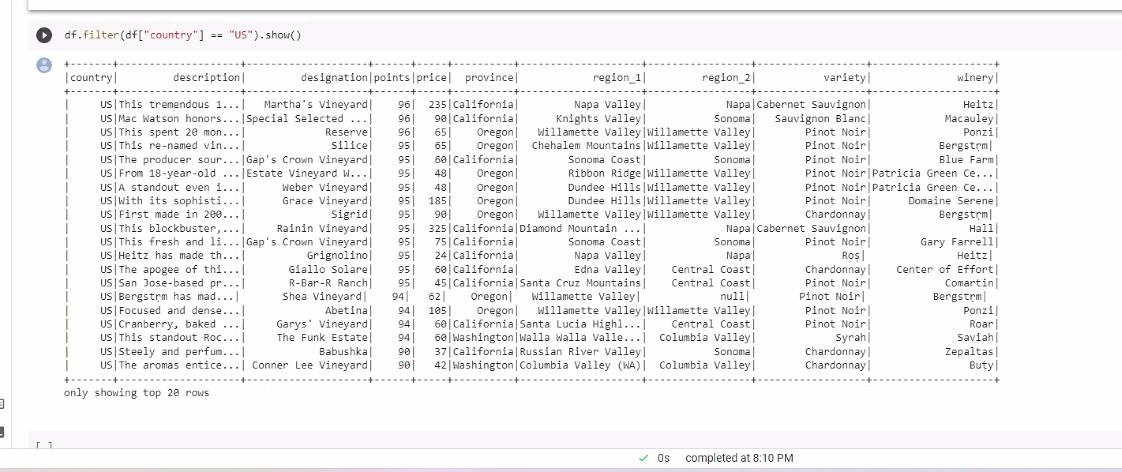


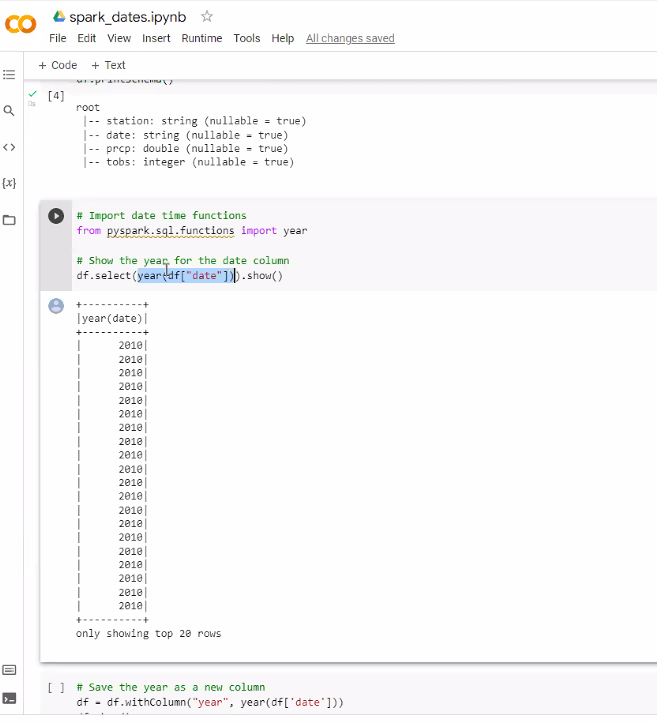


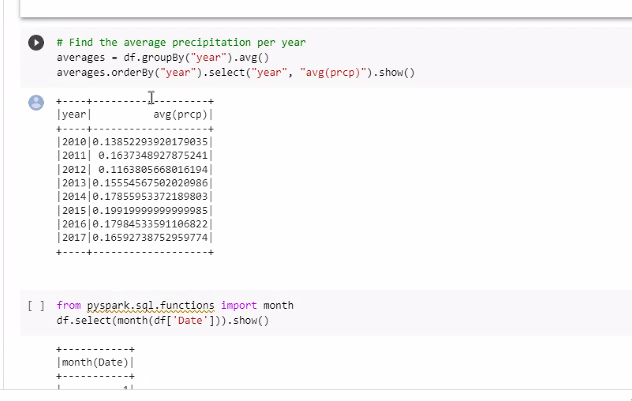




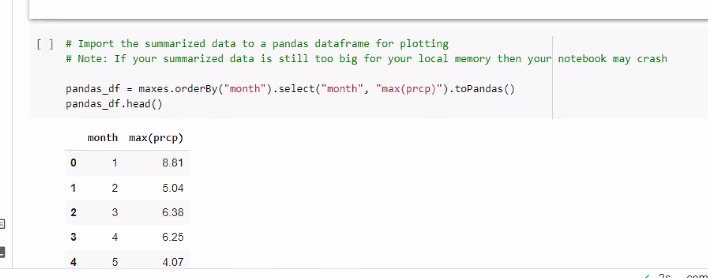
**  
**

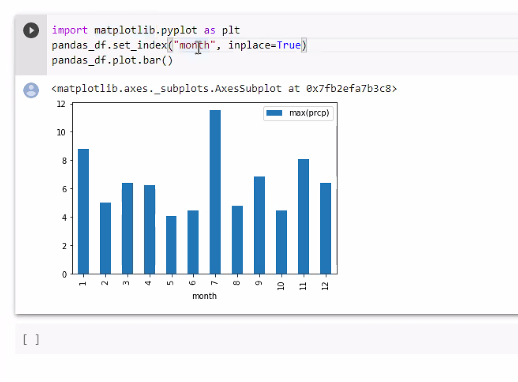
****

****

****

****

****

****

**Module 16 – start here at 16.6.2 – Stop Words**

**Module 16 – Amazon Web Services (AWS)**

AWS is the largest cloud service provider in the world. We will be creating a relational database service or RDS on the AWS.

Navigate to aws.amazon.com. Sign in to your account. Once signed in, you will be directed to the AWS Management Console. At the top of the page, click the ‘Services” drop down menu. Here you can search for RDS or scroll down and find it under the database section. Once found, click the RDS link. On the Amazon RDS page, you will see an orange button that says Create database. Go ahead and click that link. On the creation page we have two major options. We have a Standard Create which will allow us to customize every setting for our database. We also have an Easy Create which configures the database to their best practice settings and leaves us with only a few settings we need to define. We will use standard create for this demonstration. Next we will select PostgreSQL as our database engine, but note that AWS allows other engines such as MySQL and MS SQL Server. Next we will choose a template to meet our needs. Currently, our needs are not spending any money so we will select the free tier template. Depending on your usage in the future, the free tier may not be enough for your project and you may have to adjust your settings to accommodate. We will then need to name our database instance. A default identifier will automatically be fill in, but this can be changed. We will change the name to datavis. Next, we need to create a username and password for our new database. This is not the same username and password for your AWS account. This will be a new username and password specific to the database. We will keep the username postpres and crate a super secret master password. Please remember and keep this password safe as it can be annoying to reset. You also have the option to allow Amazon to automatically generate a password for you. We will then scroll down to the Connectivity section. Click the Additional connectivity configuration dropdown to display more options. The setting we need to change is under Publicly accessible. It is set to “Not” by default, but we will set it to “Yes” to be able to connect to it via Python or pgAdmin at a later time. Once finished, scroll to the bottom and click the “Create Database button.” You will then be directed back to the database page. From here you will see your newly created RDS. However, it may take a few minutes to create on Amazon’s side. Once the status changes to “Available,” you are ready to connect to your cloud database hosted by AWS.

Instructions from class

mypostgresdb.czmspp9a7s09.us-east-2.rds.amazonaws.com

my\_aws\_postgresdb\_rds

Wrong password, wrong UN, or security group was not configured correctly

