

Week 3 Module Assignment

All The Codes and Data Schema Screenshots:

The screenshot displays a Zeppelin Notebook interface within a web browser. The browser's address bar shows the URL `192.168.126.129:9995/#/notebook/2JR1RUUGE`. The Zeppelin header features the logo, the text "Notebook Job", a search bar, and a user indicator "anonymous". The notebook title is "spark_example".

The notebook contains three code blocks, each with a "FINISHED" status and execution controls:

```
%spark2.pyspark
heart_disease = spark.read.option("header",True).csv("/tmp/data/heart_disease.csv")
```

```
%spark2.pyspark
heart_disease.printSchema()

root
|-- PatientID: string (nullable = true)
|-- Age: string (nullable = true)
|-- ChestPainType: string (nullable = true)
|-- RestingBP: string (nullable = true)
|-- Cholesterol: string (nullable = true)
|-- FastingBS: string (nullable = true)
|-- RestingECG: string (nullable = true)
|-- MaxHR: string (nullable = true)
|-- ExerciseAngina: string (nullable = true)
|-- Oldpeak: string (nullable = true)
|-- ST_Slope: string (nullable = true)
|-- HeartDisease: string (nullable = true)
```

Below the second code block, a status message reads: "Took 0 sec. Last updated by anonymous at February 06 2024, 4:48:10 PM. (outdated)".


The third code block is currently empty:

```
%spark2.pyspark
```

Hortonworks Sandbox with | root@sandbox-hdp:~ - Shell | 192.168.126.129:9995/#/note | 192.168.126.129:9995/#/not | 192.168.126.129:9995/#/inte | +

← → ↺ ⚠ Not secure 192.168.126.129:9995/#/notebook/2JR1RUUGE ☆ 📄 ⬇ 📱 ⚙

★ Bookmarks 📧 Gmail 📺 YouTube 🎮 Twitch 📧 Mail - Ryan Shah -... 🌐 Word 📄 Online PDF Convert... 🦉 Owl Express

 Notebook ▾ Job

Search your Notes 🔍

anonymous ▾

spark_example ▶ ⌂ 📄 📌 📁 📂 📅 📆 📇 📈 📉 📊 📋 📌 📍 📎 📏 📐 📑 📒 📓 📔 📕 📖 📗 📘 📙 📚 📛 📜 📝 📞 📟 📠 📡 📢 📣 📤 📥 📦 📧 📨 📩 📪 📫 📬 📭 📮 📯 📰 📱 📲 📳 📴 📵 📶 📷 📸 📹 📺 📻 📼 📽 📾 📿 📠 📡 📢 📣 📤 📥 📦 📧 📨 📩 📪 📫 📬 📭 📮 📯 📰 📱 📲 📳 📴 📵 📶 📷 📸 📹 📺 📻 📼 📽 📾 📿

📄 Head 🗑

🔍 ⚙ 🔒 default ▾

%spark2.pyspark

heart_disease = spark.read.options(header='True', inferSchema='True', delimiter=',').csv("/tmp/data/heart_disease.csv")

heart_disease.printSchema()

root

|-- PatientID: integer (nullable = true)

|-- Age: integer (nullable = true)

|-- ChestPainType: string (nullable = true)

|-- RestingBP: integer (nullable = true)

|-- Cholesterol: integer (nullable = true)

|-- FastingBS: integer (nullable = true)

|-- RestingECG: string (nullable = true)

|-- MaxHR: integer (nullable = true)

|-- ExerciseAngina: string (nullable = true)

|-- Oldpeak: double (nullable = true)

|-- ST_Slope: string (nullable = true)

|-- HeartDisease: integer (nullable = true)

Finished ▶ ⌂ 📄 📌 📁 📂 📅 📆 📇 📈 📉 📊 📋 📌 📍 📎 📏 📐 📑 📒 📓 📔 📕 📖 📗 📘 📙 📚 📛 📜 📝 📞 📟 📠 📡 📢 📣 📤 📥 📦 📧 📨 📩 📪 📫 📬 📭 📮 📯 📰 📱 📲 📳 📴 📵 📶 📷 📸 📹 📺 📻 📼 📽 📾 📿

Took 4 sec. Last updated by anonymous at February 06 2024, 4:40:37 PM.

%spark2.pyspark

heart_disease.show()

Finished ▶ ⌂ 📄 📌 📁 📂 📅 📆 📇 📈 📉 📊 📋 📌 📍 📎 📏 📐 📑 📒 📓 📔 📕 📖 📗 📘 📙 📚 📛 📜 📝 📞 📟 📠 📡 📢 📣 📤 📥 📦 📧 📨 📩 📪 📫 📬 📭 📮 📯 📰 📱 📲 📳 📴 📵 📶 📷 📸 📹 📺 📻 📼 📽 📾 📿

```
%spark2.pyspark
heart_disease.show()
```

FINISHED

PatientID	Age	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
20230001	40	ATA	140	289	0	Normal	172	N	0.0	Up	0
20230002	49	NAP	160	180	0	Normal	156	N	1.0	Flat	1
20230003	37	ATA	130	283	0	ST	98	N	0.0	Up	0
20230004	48	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
20230005	54	NAP	150	195	0	Normal	122	N	0.0	Up	0
20230006	39	NAP	120	339	0	Normal	170	N	0.0	Up	0
20230007	45	ATA	130	237	0	Normal	170	N	0.0	Up	0
20230008	54	ATA	110	208	0	Normal	142	N	0.0	Up	0
20230009	37	ASY	140	207	0	Normal	130	Y	1.5	Flat	1
20230010	48	ATA	120	284	0	Normal	120	N	0.0	Up	0
20230011	37	NAP	130	211	0	Normal	142	N	0.0	Up	0
20230012	58	ATA	136	164	0	ST	99	Y	2.0	Flat	1
20230013	39	ATA	120	204	0	Normal	145	N	0.0	Up	0
20230014	49	ASY	140	234	0	Normal	140	Y	1.0	Flat	1

Took 7 sec. Last updated by anonymous at February 06 2024, 4:50:44 PM.

```
%spark2.pyspark
```

FINISHED

```
%spark2.pyspark
heart_disease.count()
```

FINISHED

918

Took 1 sec. Last updated by anonymous at February 06 2024, 4:51:40 PM.

```
%spark2.pyspark
heart_disease.first()
```

FINISHED

Row(PatientID=20230001, Age=40, ChestPainType=u'ATA', RestingBP=140, Cholesterol=289, FastingBS=0, RestingECG=u'Normal', MaxHR=172, ExerciseAngina=u'N', Oldpeak=0.0, ST_Slope=u'Up', HeartDisease=0)

Took 1 sec. Last updated by anonymous at February 06 2024, 4:52:08 PM.

```
%spark2.pyspark
spark.read.options(header='True', inferSchema='True', delimiter=',').csv("/tmp/data/heart_disease.csv").createOrReplaceTempView("PATIENTS")
```

FINISHED

Took 3 sec. Last updated by anonymous at February 06 2024, 4:54:05 PM.

192.168.126.129:9995/#/notebook/2JR1RUUGE

Not secure

★ Bookmarks

Gmail

YouTube

Twitch

Mail - Ryan Shah ~...

Word

Online PDF Convert...

Owl Express

Zeppelin

Notebook

Job

Search your Notes

Q

anonymous

spark_example

▶

⌵

📄

🔗

📁

🔍

⌵

Head

🗑

default

%spark2.pyspark

spark.sql("SELECT * FROM PATIENTS").show()

FINISHED ▶ ⌵ 📄 ⚙

PatientID	Age	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
20230001	40	ATA	140	289	0	Normal	172	N	0.0	Up	0
20230002	49	NAP	160	180	0	Normal	156	N	1.0	Flat	1
20230003	37	ATA	130	283	0	ST	98	N	0.0	Up	0
20230004	48	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
20230005	54	NAP	150	195	0	Normal	122	N	0.0	Up	0
20230006	39	NAP	120	339	0	Normal	170	N	0.0	Up	0
20230007	45	ATA	130	237	0	Normal	170	N	0.0	Up	0
20230008	54	ATA	110	208	0	Normal	142	N	0.0	Up	0
20230009	37	ASY	140	207	0	Normal	130	Y	1.5	Flat	1
20230010	48	ATA	120	284	0	Normal	120	N	0.0	Up	0
20230011	37	NAP	130	211	0	Normal	142	N	0.0	Up	0
20230012	58	ATA	136	164	0	ST	99	Y	2.0	Flat	1
20230013	39	ATA	120	204	0	Normal	145	N	0.0	Up	0
20230014	49	ASY	140	234	0	Normal	140	Y	1.0	Flat	1

Took 7 sec. Last updated by anonymous at February 08 2024, 4:54:41 PM.

%spark2.pyspark

FINISHED ▶ ⌵ 📄 ⚙

[illegible]