

Applied Data Science Lab

My Path Module | WorldQuant U

work/ds-curricu - JupyterLab

Data-Science-Lab/model.py at m

(45) YouTube

vm.wqu.edu/lab/tree/work/ds-curriculum/060-consumer-finances-in-usa/065-assignment.ipynb

File Edit View Run Kernel Tabs Settings Help

Filter files by name

Name

Last Modified

065-assignment.ipynb

2 hours ago

data

2 months ago

images

seconds ago

061-explorin...

2 months ago

062-clusteri...

2 months ago

063-clusteri...

2 months ago

064-interact...

2 months ago

065-assign...

seconds ago

066-data-di...

2 months ago

6.5. Small Business Owners in the United States

In this assignment, you're going to focus on business owners in the United States. You'll start by examining some demographic characteristics of the group, such as age, income category, and debt vs home value. Then you'll select high-variance features, and create a clustering model to divide small business owners into subgroups. Finally, you'll create some visualizations to highlight the differences between these subgroups. Good luck! 🍀

```
[1]: import wqet_grader
wqet_grader.init("Project 6 Assessment")
```

```
[4]: # Import libraries here
import matplotlib.pyplot as plt
import plotly.express as px
import pandas as pd
import seaborn as sns
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score
from teaching_tools.widgets import ClusterWidget, SCFClusterWidget
from scipy.stats.mstats import trimmed_var
from sklearn.decomposition import PCA
from sklearn.pipeline import make_pipeline
from sklearn.preprocessing import StandardScaler
```

Prepare Data

Import

Let's start by bringing our data into the assignment.

Task 6.5.1: Read the file "data/SCFP2019.csv.gz" into the DataFrame `df`.

Simple 0 1 Python 3 (pykernel) | Idle Mode: Command Ln 1, Col 1 English (United States) 065-assignment.ipynb 30°C Haze Search ENG IN 18:44 18-03-2023

Applied Data Science Lab

My Path Module | WorldQuant U

work/ds-curricu - JupyterLab

Data-Science-Lab/model.py at m

(45) YouTube

vm.wqu.edu/lab/tree/work/ds-curriculum/060-consumer-finances-in-usa/065-assignment.ipynb

File Edit View Run Kernel Tabs Settings Help

Filter files by name

Name

Last Modified

065-assignment.ipynb

2 hours ago

data

2 months ago

images

seconds ago

061-explorin...

2 months ago

062-clusteri...

2 months ago

063-clusteri...

2 months ago

064-interact...

2 months ago

065-assign...

seconds ago

066-data-di...

2 months ago

Prepare Data

Import

Let's start by bringing our data into the assignment.

Task 6.5.1: Read the file "data/SCFP2019.csv.gz" into the DataFrame `df`.

```
[5]: df = pd.read_csv("data/SCFP2019.csv.gz")
print("df shape:", df.shape)
df.head()
```

```
df shape: (28885, 351)
```

```
[5]:
```

	YY1	YY1	WGT	HHSEX	AGE	AGECL	EDUC	EDCL	MARRIED	KIDS	...	NWCAT	INCCAT	ASSETCAT	NINCCAT	NINC2CAT	NWPC1LECAT	INCP1LECAT	NINCP1LECAT	INCQRTCAT	NINQRTCAT
0	1	11	6119.779308	2	75	6	12	4	2	0	...	5	3	6	3	2	10	6	6	3	
1	1	12	4712.374912	2	75	6	12	4	2	0	...	5	3	6	3	1	10	5	5	2	
2	1	13	5145.224455	2	75	6	12	4	2	0	...	5	3	6	3	1	10	5	5	2	
3	1	14	5297.663412	2	75	6	12	4	2	0	...	5	2	6	2	1	10	4	4	2	
4	1	15	4761.812371	2	75	6	12	4	2	0	...	5	3	6	3	1	10	5	5	2	

5 rows × 351 columns

```
[6]: wqet_grader.grade("Project 6 Assessment", "Task 6.5.1", list(df.shape))
```

✓

Yup. You got it.
Score: 1

Simple 0 1 Python 3 (pykernel) | Idle Mode: Command Ln 1, Col 1 English (United States) 065-assignment.ipynb 30°C Haze Search ENG IN 18:44 18-03-2023

Applied Data Science Lab | My Path Module | WorldQuant U | work/ds-curricu - JupyterLab | Data-Science-Lab/model.py at n | (45) YouTube

vm.wqu.edu/lab/tree/work/ds-curriculum/060-consumer-finances-in-usa/065-assignment.ipynb

File Edit View Run Kernel Tabs Settings Help

065-assignment.ipynb

Filter files by name

Name	Last Modified
065-assignment.ipynb	2 hours ago
data	2 months ago
images	seconds ago
061-explorin...	2 months ago
062-clusteri...	2 months ago
063-clusteri...	2 months ago
064-interact...	2 months ago
065-assign...	seconds ago
066-data-di...	2 months ago

Explore

As mentioned at the start of this assignment, you're focusing on business owners. But what percentage of the respondents in `df` are business owners?

Task 6.5.2: Calculate the proportion of respondents in `df` that are business owners, and assign the result to the variable `pct_biz_owners`. You'll need to review the documentation regarding the `HBUS` column to complete these tasks.

```
[13]: prop_biz_owners = sum(df["HBUS"]) / (sum(df["HBUS"] == 0) + sum(df["HBUS"]))
print("% of business owners in df:", prop_biz_owners)
```

% of business owners in df: 0.2740176562229531

```
[14]: wqet_grader.grade("Project 6 Assessment", "Task 6.5.2", [prop_biz_owners])
```

That's the right answer. Keep it up!
Score: 1

Is the distribution of income different for business owners and non-business owners?

Task 6.5.3: Create a DataFrame `df_inccat` that shows the normalized frequency for income categories for business owners and non-business owners. Your final DataFrame should look something like this:

	HBUS	INCCAT	frequency
0	0	0-20	0.210348
1	0	21-39.9	0.198140
...			
11	1	0-20	0.041188

```
[15]: inccat_dict = {
1: "0-20",
2: "21-39.9",
3: "40-59.9",
4: "60-79.9",
5: "80-99.9",
6: "100-119.9"
}
```

Simple | 0 | 1 | Python 3 (pykernel) | Idle

Mode: Command | Ln 1, Col 1 | English (United States) | 065-assignment.ipynb

30°C
Haze

Search

Applied Data Science Lab | My Path Module | WorldQuant U | work/ds-curricu - JupyterLab | Data-Science-Lab/model.py at n | (45) YouTube

vm.wqu.edu/lab/tree/work/ds-curriculum/060-consumer-finances-in-usa/065-assignment.ipynb

File Edit View Run Kernel Tabs Settings Help

065-assignment.ipynb

Filter files by name

Name	Last Modified
065-assignment.ipynb	2 hours ago
data	2 months ago
images	seconds ago
061-explorin...	2 months ago
062-clusteri...	2 months ago
063-clusteri...	2 months ago
064-interact...	2 months ago
065-assign...	seconds ago
066-data-di...	2 months ago

Is the distribution of income different for business owners and non-business owners?

Task 6.5.3: Create a DataFrame `df_inccat` that shows the normalized frequency for income categories for business owners and non-business owners. Your final DataFrame should look something like this:

	HBUS	INCCAT	frequency
0	0	0-20	0.210348
1	0	21-39.9	0.198140
...			
11	1	0-20	0.041188

```
[15]: inccat_dict = {
1: "0-20",
2: "21-39.9",
3: "40-59.9",
4: "60-79.9",
5: "80-99.9",
6: "100-119.9"
}

df_inccat = (
df["INCCAT"]
.replace(inccat_dict)
.groupby(df["HBUS"])
.value_counts(normalize=True)
.rename("frequency")
.to_frame()
.reset_index()
)

df_inccat
```

	HBUS	INCCAT	frequency
0	0	0-20	0.210348
1	0	21-39.9	0.198140
2	0	40-59.9	0.189080
3	0	60-79.9	0.186600

Simple | 0 | 1 | Python 3 (pykernel) | Idle

Mode: Command | Ln 1, Col 1 | English (United States) | 065-assignment.ipynb

30°C
Haze

Search

Applied Data Science Lab | My Path Module | WorldQuant U | work/ds-curricu - JupyterLab | Data-Science-Lab/model.py at n | (45) YouTube

Applied Data Science Lab | My Path Module | WorldQuant U | work/ds-curricu - JupyterLab | Data-Science-Lab/model.py at n | (45) YouTube

vm.wqu.edu/lab/tree/work/ds-curriculum/060-consumer-finances-in-usa/065-assignment.ipynb

File Edit View Run Kernel Tabs Settings Help

065-assignment.ipynb

Filter files by name

Name	Last Modified
065-assignment.ipynb	2 hours ago
data	2 months ago
images	seconds ago
061-explorin...	2 months ago
062-clusteri...	2 months ago
063-clusteri...	2 months ago
064-interact...	2 months ago
065-assign...	seconds ago
066-data-di...	2 months ago

Is the distribution of income different for business owners and non-business owners?

Task 6.5.3: Create a DataFrame `df_inccat` that shows the normalized frequency for income categories for business owners and non-business owners. Your final DataFrame should look something like this:

	HBUS	INCCAT	frequency
0	0	0-20	0.210348
1	0	21-39.9	0.198140
...			
11	1	0-20	0.041188

```
[15]: inccat_dict = {
1: "0-20",
2: "21-39.9",
3: "40-59.9",
4: "60-79.9",
5: "80-99.9",
6: "100-119.9"
}

df_inccat = (
df["INCCAT"]
.replace(inccat_dict)
.groupby(df["HBUS"])
.value_counts(normalize=True)
.rename("frequency")
.to_frame()
.reset_index()
)

df_inccat
```

	HBUS	INCCAT	frequency
0	0	0-20	0.210348
1	0	21-39.9	0.198140
2	0	40-59.9	0.189080
3	0	60-79.9	0.186600

Simple | 0 | 1 | Python 3 (pykernel) | Idle

Mode: Command | Ln 1, Col 1 | English (United States) | 065-assignment.ipynb

30°C
Haze

Search

Applied Data Science Lab | My Path Module | WorldQuant U | work/ds-curricu - JupyterLab | Data-Science-Lab/model.py at n | (45) YouTube

Applied Data Science Lab | My Path Module | WorldQuant U | work/ds-curricu - JupyterLab | Data-Science-Lab/model.py at n | (45) YouTube

vm.wqu.edu/lab/tree/work/ds-curriculum/060-consumer-finances-in-usa/065-assignment.ipynb

File Edit View Run Kernel Tabs Settings Help

065-assignment.ipynb

Filter files by name

/ / ds-curriculum / 060-consumer-finances-in-usa /

Name	Last Modified
065-assignment.ipynb	seconds ago
066-data-di...	2 months ago
061-explorin...	2 months ago
062-clusteri...	2 months ago
063-clusteri...	2 months ago
064-interact...	2 months ago
065-assignment.ipynb	seconds ago
066-data-di...	2 months ago

df_inccat

```
[15]:
```

	HBUS	INCCAT	frequency
0	0	0-20	0.210348
1	0	21-39.9	0.198140
2	0	40-59.9	0.189080
3	0	60-79.9	0.186600
4	0	80-90	0.117167
5	0	80-89.9	0.090665
6	1	90-100	0.629438
7	1	60-79.9	0.119015
8	1	80-89.9	0.097410
9	1	40-59.9	0.071510
10	1	21-39.9	0.041440
11	1	0-20	0.041188

```
[16]: wqet_grader.grade("Project 6 Assessment", "Task 6.5.3", df_inccat)
```

✓ You got it, Dance party time! 🎉 🎉 🎉
Score: 1

Task 6.5.4: Using seaborn, create a side-by-side bar chart of df_inccat. Set hue to "HBUS", and make sure that the income categories are in the correct order along the x-axis. Label to the x-axis "Income Category", the y-axis "Frequency (%)", and use the title "Income Distribution: Business Owners vs. Non-Business Owners".

```
[17]: # Create bar chart of 'df_inccat'
```

```
sns.barplot(
    xx="INCCAT",
    y="frequency",
    hue="HBUS",
    data=df_inccat,
    order=df_inccat.INCCAT.unique()
)
plt.xlabel("<Your x_title>")
plt.ylabel("<Your y_title>")
plt.title("<Your Title>");
```

Don't delete the code below 🌟
plt.savefig("images/6-5-4.png", dpi=150)

```
[18]: with open("images/6-5-4.png", "rb") as file:
    wqet_grader.grade("Project 6 Assessment", "Task 6.5.4", file)
```

Simple 0 1 Python 3 (ipykernel) | Idle

Mode: Command Ln 1, Col 1 English (United States) 065-assignment.ipynb

30°C
Haze

Search

Applied Data Science Lab | My Path Module | WorldQuant U | work/ds-curricu - JupyterLab | Data-Science-Lab/model.py at n | (45) YouTube

vm.wqu.edu/lab/tree/work/ds-curriculum/060-consumer-finances-in-usa/065-assignment.ipynb

File Edit View Run Kernel Tabs Settings Help

065-assignment.ipynb

Filter files by name

/ / ds-curriculum / 060-consumer-finances-in-usa /

Name	Last Modified
065-assignment.ipynb	seconds ago
066-data-di...	2 months ago
061-explorin...	2 months ago
062-clusteri...	2 months ago
063-clusteri...	2 months ago
064-interact...	2 months ago
065-assignment.ipynb	seconds ago
066-data-di...	2 months ago

Task 6.5.4: Using seaborn, create a side-by-side bar chart of df_inccat. Set hue to "HBUS", and make sure that the income categories are in the correct order along the x-axis. Label to the x-axis "Income Category", the y-axis "Frequency (%)", and use the title "Income Distribution: Business Owners vs. Non-Business Owners".

```
[17]: # Create bar chart of 'df_inccat'
```

```
sns.barplot(
    xx="INCCAT",
    y="frequency",
    hue="HBUS",
    data=df_inccat,
    order=df_inccat.INCCAT.unique()
)
plt.xlabel("<Your x_title>")
plt.ylabel("<Your y_title>")
plt.title("<Your Title>");
```

Don't delete the code below 🌟
plt.savefig("images/6-5-4.png", dpi=150)

```
[18]: with open("images/6-5-4.png", "rb") as file:
    wqet_grader.grade("Project 6 Assessment", "Task 6.5.4", file)
```

Simple 0 1 Python 3 (ipykernel) | Idle

Mode: Command Ln 1, Col 1 English (United States) 065-assignment.ipynb

30°C
Haze

Search

Applied Data Science Lab | My Path Module | WorldQuant U | work/ds-curricu - JupyterLab | Data-Science-Lab/model.py at n | (45) YouTube

Applied Data Science Lab | My Path Module | WorldQuant U | work/ds-curriculum - JupyterLab | Data-Science-Lab/model.py at m | (45) YouTube

vm.wqu.edu/lab/tree/work/ds-curriculum/060-consumer-finances-in-usa/065-assignment.ipynb

File Edit View Run Kernel Tabs Settings Help

065-assignment.ipynb

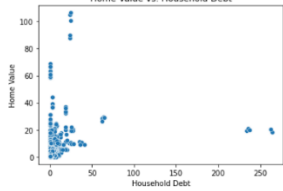
Filter files by name

Name Last Modified

- ipynb_chec... 3 hours ago
- data 2 months ago
- images a minute ago
- 061-explorin... 2 months ago
- 062-clusteri... 2 months ago
- 063-clusteri... 2 months ago
- 064-interact... 2 months ago
- 065-assign... a minute ago
- 066-data-di... 2 months ago

Task 6.5.5: Using seaborn, create a scatter plot that shows "HOUSESES" vs. "DEBT". You should color the datapoints according to business ownership. Be sure to label the x-axis "Household Debt", the y-axis "Home Value", and use the title "Home Value vs. Household Debt".

```
[19]: # Plot "HOUSESES" vs "DEBT" with hue=Label
sns.scatterplot(x=df["DEBT"] / 1e6, y=df["HOUSESES"] / 1e6, palette="deep")
plt.xlabel("Household Debt")
plt.ylabel("Home Value")
plt.title("Home Value vs. Household Debt");
# Don't delete the code below
plt.savefig("images/6-5-5.png", dpi=150)
```



For the model building part of the assignment, you're going to focus on small business owners, defined as respondents who have a business and whose income does not exceed \$500,000.

```
[21]: with open("images/6-5-5.png", "rb") as file:
      wget_grader.grade("Project 6 Assessment", "Task 6.5.5", file)
```

Yes! Keep on rockin'. 🎸 That's right.
Score: 1

Simple 0 1 Python 3 (ipykernel) | Idle Mode: Command Ln 1, Col 1 English (United States) 065-assignment.ipynb

30°C Haze Search

Applied Data Science Lab | My Path Module | WorldQuant U | work/ds-curriculum - JupyterLab | Data-Science-Lab/model.py at m | (45) YouTube

vm.wqu.edu/lab/tree/work/ds-curriculum/060-consumer-finances-in-usa/065-assignment.ipynb

File Edit View Run Kernel Tabs Settings Help

065-assignment.ipynb

Filter files by name

Name Last Modified

- ipynb_chec... 3 hours ago
- data 2 months ago
- images a minute ago
- 061-explorin... 2 months ago
- 062-clusteri... 2 months ago
- 063-clusteri... 2 months ago
- 064-interact... 2 months ago
- 065-assign... a minute ago
- 066-data-di... 2 months ago

Task 6.5.6: Create a new DataFrame `df_small_biz` that contains only business owners whose income is below \$500,000.

```
[22]: mask = (df["HBUS"]) & (df["INCOME"] < 500_000)
df_small_biz = df[mask]
print(df_small_biz.shape, df_small_biz.shape)
df_small_biz.head()
```

```
[22]: YY1 YY1 WGT HHSEX AGE AGECL EDUC EDCL MARRIED KIDS ... NWCAT INCCAT ASSETCAT NINCCAT NINC2CAT NHPCTLECAT INCPCTLECAT NINCPCTLECAT INCORTCAT NINQRTK
80 17 171 7802.265717 1 62 4 12 4 1 0 ... 3 5 5 5 2 7 9 9 4
81 17 172 8247.536301 1 62 4 12 4 1 0 ... 3 5 5 5 2 7 9 9 4
82 17 173 8169.562719 1 62 4 12 4 1 0 ... 3 5 5 5 2 7 9 9 4
83 17 174 8087.704517 1 62 4 12 4 1 0 ... 3 5 5 5 2 7 9 9 4
84 17 175 8276.510048 1 62 4 12 4 1 0 ... 3 5 5 5 2 7 9 9 4
```

5 rows × 351 columns

```
[23]: wget_grader.grade("Project 6 Assessment", "Task 6.5.6", list(df_small_biz.shape))
```

Way to go!
Score: 1

We saw that credit-fearful respondents were relatively young. Is the same true for small business owners?

Task 6.5.7: Create a histogram from the "AGE" column in `df_small_biz` with 10 bins. Be sure to label the x-axis "Age", the y-axis "Frequency (count)", and use the title "Small Business Owners: Age Distribution".

```
[24]: # Plot histogram of "AGE"
df_small_biz["AGE"].hist(bins=10)
plt.xlabel("Your x_label")
plt.ylabel("Your y_label")
```

Simple 0 1 Python 3 (ipykernel) | Idle Mode: Command Ln 1, Col 1 English (United States) 065-assignment.ipynb

30°C Haze Search

Applied Data Science Lab | My Path Module | WorldQuant U | work/ds-curricu - JupyterLab | Data-Science-Lab/model.py at n | (45) YouTube

vm.wqu.edu/lab/tree/work/ds-curriculum/060-consumer-finances-in-usa/065-assignment.ipynb

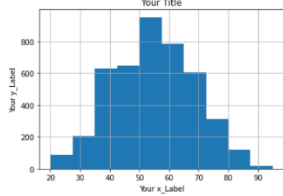
File Edit View Run Kernel Tabs Settings Help

065-assignment.ipynb

We saw that credit-fearful respondents were relatively young. Is the same true for small business owners?

Task 6.5.7: Create a histogram from the "AGE" column in `df_small_biz` with 10 bins. Be sure to label the x-axis "Age", the y-axis "Frequency (count)", and use the title "Small Business Owners: Age Distribution".

```
[24]: # Plot histogram of "AGE"
df_small_biz["AGE"].hist(bins=10)
plt.xlabel("Your x_label")
plt.ylabel("Your y_label")
plt.title("Your Title");
# Don't delete the code below
plt.savefig("images/6-5-7.png", dpi=150)
```



So, can we say the same thing about small business owners as we can about credit-fearful people?

```
[25]: with open("images/6-5-7.png", "rb") as file:
wqet_grader.grade("Project 6 Assessment", "Task 6.5.7", file)
```

Yes! Great problem solving.
Score: 1

Single | 0 | 1 | Python 3 (ipykernel) | Idle

Mode: Command | Ln 1, Col 1 | English (United States) | 065-assignment.ipynb

30°C Haze | Search | ENG IN | 18:45 18-03-2023

Applied Data Science Lab | My Path Module | WorldQuant U | work/ds-curricu - JupyterLab | Data-Science-Lab/model.py at n | (45) YouTube

vm.wqu.edu/lab/tree/work/ds-curriculum/060-consumer-finances-in-usa/065-assignment.ipynb

File Edit View Run Kernel Tabs Settings Help

065-assignment.ipynb

Task 6.5.8: Calculate the variance for all the features in `df_small_biz`, and create a Series `top_ten_var` with the 10 features with the largest variance.

```
[26]: # Calculate variance, get 10 largest features
top_ten_var = df_small_biz.var().sort_values().tail(10)
top_ten_var
```

```
[26]: EQUITY    1.005088e+13
FIN        2.103228e+13
KGBUS      5.025218e+13
ACTBUS      5.405021e+13
BUS         5.606717e+13
KGTOTAL     6.120760e+13
NHWFIN      7.363197e+13
NFIN        9.244074e+13
NETWORTH    1.424450e+14
ASSET       1.520071e+14
dtype: float64
```

```
[27]: wqet_grader.grade("Project 6 Assessment", "Task 6.5.8", top_ten_var)
```

Yes! Great problem solving.
Score: 1

We'll need to remove some outliers to avoid problems in our calculations, so let's trim them out.

Task 6.5.9: Calculate the trimmed variance for the features in `df_small_biz`. Your calculations should not include the top and bottom 10% of observations. Then create a Series `top_ten_trim_var` with the 10 features with the largest variance.

```
[28]: # Calculate trimmed variance
top_ten_trim_var = df_small_biz.apply(trimmed_var, limits=(0.1, 0.1)).sort_values().tail(10)
top_ten_trim_var
```

```
[28]: EQUITY    1.177020e+11
KGBUS      1.838163e+11
FIN        3.588855e+11
KGTOTAL     6.120760e+11
```

Single | 0 | 1 | Python 3 (ipykernel) | Idle

Mode: Command | Ln 1, Col 1 | English (United States) | 065-assignment.ipynb

30°C Haze | Search | ENG IN | 18:45 18-03-2023

Applied Data Science Lab | My Path Module | WorldQuant | work/ds-curricu - JupyterLab | Data-Science-Lab/model.py at n | (45) YouTube

vm.wqu.edu/lab/tree/work/ds-curriculum/060-consumer-finances-in-usa/065-assignment.ipynb

File Edit View Run Kernel Tabs Settings Help

065-assignment.ipynb

Filter files by name

Name	Last Modified
065-assignment.ipynb	3 hours ago
data	2 months ago
images	a minute ago
061-explorin...	2 months ago
062-cluster...	2 months ago
063-cluster...	2 months ago
064-interact...	2 months ago
065-assign...	a minute ago
066-data-di...	2 months ago

Task 6.5.9: Calculate the trimmed variance for the features in `df_small_biz`. Your calculations should not include the top and bottom 10% of observations. Then create a Series `top_ten_trim_var` with the 10 features with the largest variance.

```
[28]: # Calculate trimmed variance
top_ten_trim_var = df_small_biz.apply(trimmed_var, limits=(0.1, 0.1)).sort_values().tail(10)
top_ten_trim_var
```

```
[29]: wqet_grader.grade("Project 6 Assessment", "Task 6.5.9", top_ten_trim_var)
```

Python master 🏆
Score: 1

Let's do a quick visualization of those values.

Task 6.5.10: Use `plotly` express to create a horizontal bar chart of `top_ten_trim_var`. Be sure to label your x-axis "Trimmed Variance [\$]", the y-axis "Feature", and use the title "Small Business Owners: High Variance Features".

```
[33]: # Create horizontal bar chart of 'top_ten_trim_var'
fig = px.bar(
    x=top_ten_trim_var,
    y=top_ten_trim_var.index,
    title="High Var Feat"
)
fig.update_layout(xaxis_title="Your x_label", yaxis_title="Your y_label")
# Don't delete the code below
fig.write_image("images/6-5-10.png", scale=1, height=500, width=700)
fig.show()
```

30°C
Haze

Applied Data Science Lab | My Path Module | WorldQuant | work/ds-curricu - JupyterLab | Data-Science-Lab/model.py at n | (45) YouTube

vm.wqu.edu/lab/tree/work/ds-curriculum/060-consumer-finances-in-usa/065-assignment.ipynb

File Edit View Run Kernel Tabs Settings Help

065-assignment.ipynb

Filter files by name

Name	Last Modified
065-assignment.ipynb	3 hours ago
data	2 months ago
images	a minute ago
061-explorin...	2 months ago
062-cluster...	2 months ago
063-cluster...	2 months ago
064-interact...	2 months ago
065-assign...	a minute ago
066-data-di...	2 months ago

Let's do a quick visualization of those values.

Task 6.5.10: Use `plotly` express to create a horizontal bar chart of `top_ten_trim_var`. Be sure to label your x-axis "Trimmed Variance [\$]", the y-axis "Feature", and use the title "Small Business Owners: High Variance Features".

```
[33]: # Create horizontal bar chart of 'top_ten_trim_var'
fig = px.bar(
    x=top_ten_trim_var,
    y=top_ten_trim_var.index,
    title="High Var Feat"
)
fig.update_layout(xaxis_title="Your x_label", yaxis_title="Your y_label")
# Don't delete the code below
fig.write_image("images/6-5-10.png", scale=1, height=500, width=700)
fig.show()
```

High Var Feat

```
[34]: with open("images/6-5-10.png", "rb") as file:
wqet_grader.grade("Project 6 Assessment", "Task 6.5.10", file)
```

30°C
Haze

Applied Data Science Lab | My Path Module | WorldQuant U | work/ds-curricu - JupyterLab | Data-Science-Lab/model.py at m | (45) YouTube

vm.wqu.edu/lab/tree/work/ds-curriculum/060-consumer-finances-in-usa/065-assignment.ipynb

File Edit View Run Kernel Tabs Settings Help

065-assignment.ipynb

Markdown Python 3 (pykernel)

Based on this graph, which five features have the highest variance?

Task 6.5.11: Generate a list `high_var_cols` with the column names of the five features with the highest trimmed variance.

```
[35]: high_var_cols = top_ten_trim_var.tail(5).index.to_list()
high_var_cols
```

```
[35]: ['BUS', 'HHFIN', 'NFIN', 'NETWORTH', 'ASSET']
```

```
[36]: wqet_grader.grade("Project 6 Assessment", "Task 6.5.11", high_var_cols)
```

✓ You are coding
Score: 1

Split

Let's turn that list into a feature matrix.

Task 6.5.12: Create the feature matrix `X`. It should contain the five columns in `high_var_cols`.

```
[37]: X = df_small_biz[high_var_cols]
print("X shape:", X.shape)
X shape: (4364, 5)
```

```
[38]: wqet_grader.grade("Project 6 Assessment", "Task 6.5.12", list(X.shape))
```

✓ Correct.
Score: 1

Build Model

Simple Python 3 (pykernel) | Idle Mode: Command Ln 1, Col 1 English (United States) 065-assignment.ipynb

30°C Haze

Applied Data Science Lab | My Path Module | WorldQuant U | work/ds-curricu - JupyterLab | Data-Science-Lab/model.py at m | (45) YouTube

vm.wqu.edu/lab/tree/work/ds-curriculum/060-consumer-finances-in-usa/065-assignment.ipynb

File Edit View Run Kernel Tabs Settings Help

065-assignment.ipynb

Markdown Python 3 (pykernel)

Build Model

Now that our data is in order, let's get to work on the model.

Iterate

Task 6.5.13: Use a `for` loop to build and train a K-Means model where `n_clusters` ranges from 2 to 12 (inclusive). Your model should include a `StandardScaler`. Each time a model is trained, calculate the inertia and add it to the list `inertia_errors`, then calculate the silhouette score and add it to the list `silhouette_scores`.

Note: For reproducibility, make sure you set the random state for your model to `42`.

```
[39]: n_clusters = range(2, 13)
inertia_errors = []
silhouette_scores = []

# Use for loop
for k in n_clusters:
    # Build
    model = make_pipeline(StandardScaler(), KMeans(n_clusters=k, random_state=42))
    # Train
    model.fit(X)
    # Calculate inertia
    inertia_errors.append(model.named_steps["kmeans"].inertia_)
    # Calculate silhouette score
    silhouette_scores.append(
        silhouette_score(X, model.named_steps["kmeans"].labels_)
    )

print("Inertia:", inertia_errors[:10])
print()
print("Silhouette Scores:", silhouette_scores[:3])

Inertia: [5765.863949365048, 3070.4294488357455, 2220.292185089684, 1777.4635570665569, 1443.7860071034045, 1173.3701169574997, 1004.0082329287382, 892.7197264630449, 780.7646441851751, 678.9317940468646]
```


Applied Data Science Lab | My Path Module | WorldQuant | work/ds-curriculum - JupyterLab | Data-Science-Lab/model.py at | (45) YouTube

vm.wqu.edu/lab/tree/work/ds-curriculum/060-consumer-finances-in-usa/065-assignment.ipynb

File Edit View Run Kernel Tabs Settings Help

065-assignment.ipynb

Filter files by name

Name	Last Modified
.ipynb_checkpoints	3 hours ago
data	2 months ago
images	2 minutes ago
061-explorin...	2 months ago
062-clusteri...	2 months ago
063-clusteri...	2 months ago
064-interact...	2 months ago
065-assign...	2 minutes ago
066-data-di...	2 months ago

```

Inertia: [5765.863949365848, 3070.4294488357455, 2220.292185089684, 1777.4635570665569, 1443.7868071034045, 1173.3701169574997, 1004.0082329287382, 892.7197264630449, 780.764641851751, 678.9317940468646]

Silhouette Scores: [0.9542706303253067, 0.8446503900103915, 0.7422220122162623]

[40]: wqet_grader.grade("Project 6 Assessment", "Task 6.5.13", list(inertia_errors))

Wow, you're making great progress.
Score: 1

Just like we did in the previous module, we can start to figure out how many clusters we'll need with a line plot based on inertia.

Task 6.5.14: Use plotly express to create a line plot that shows the values of inertia_errors as a function of n_clusters. Be sure to label your x-axis "Number of Clusters", your y-axis "Inertia", and use the title "K-Means Model: Inertia vs Number of Clusters".

[41]: # Create line plot of 'inertia_errors' vs 'n_clusters'
fig = px.line(
    x=n_clusters, y=inertia_errors, title="Your Title"
)
fig.update_layout(xaxis_title="Your x_label", yaxis_title="Your y_label")
fig.write_image("images/6-5-14.png", scale=1, height=500, width=700)
fig.show()

Your Title

6000
4000
2000
0
2 4 6 8 10 12
Your x_label
Your y_label

```

Simple 0 1 Python 3 (ipykernel) | Idle Mode: Command Ln 1, Col 1 English (United States) 065-assignment.ipynb

30°C Haze Search ENG IN 18:45 18-03-2023

Applied Data Science Lab | My Path Module | WorldQuant | work/ds-curriculum - JupyterLab | Data-Science-Lab/model.py at | (45) YouTube

vm.wqu.edu/lab/tree/work/ds-curriculum/060-consumer-finances-in-usa/065-assignment.ipynb

File Edit View Run Kernel Tabs Settings Help

065-assignment.ipynb

Filter files by name

Name	Last Modified
.ipynb_checkpoints	3 hours ago
data	2 months ago
images	2 minutes ago
061-explorin...	2 months ago
062-clusteri...	2 months ago
063-clusteri...	2 months ago
064-interact...	2 months ago
065-assign...	2 minutes ago
066-data-di...	2 months ago

```

[42]: with open("images/6-5-14.png", "rb") as file:
wqet_grader.grade("Project 6 Assessment", "Task 6.5.14", file)

Good work!
Score: 1

And let's do the same thing with our Silhouette Scores.

Task 6.5.15: Use plotly express to create a line plot that shows the values of silhouette_scores as a function of n_clusters. Be sure to label your x-axis "Number of Clusters", your y-axis "Silhouette Score", and use the title "K-Means Model: Silhouette Score vs Number of Clusters".

[43]: # Create a line plot of 'silhouette_scores' vs 'n_clusters'
fig = px.line(
    x=n_clusters, y=silhouette_scores, title="Your Title"
)
fig.update_layout(xaxis_title="Your x_label", yaxis_title="Your y_label")
fig.write_image("images/6-5-15.png", scale=1, height=500, width=700)
fig.show()

Your Title

6000
4000
2000
0
2 4 6 8 10 12
Your x_label
Your y_label

```

Simple 0 1 Python 3 (ipykernel) | Idle Mode: Command Ln 1, Col 1 English (United States) 065-assignment.ipynb

30°C Haze Search ENG IN 18:45 18-03-2023

Applied Data Science Lab | My Path Module | WorldQuant | work/ds-curricu - JupyterLab | Data-Science-Lab/model.py at n | (45) YouTube

vm.wqu.edu/lab/tree/work/ds-curriculum/060-consumer-finances-in-usa/065-assignment.ipynb

File Edit View Run Kernel Tabs Settings Help

065-assignment.ipynb

Filter files by name

Name	Last Modified
065-assignment.ipynb	3 hours ago
data	2 months ago
images	2 minutes ago
061-explorin...	2 months ago
062-clusteri...	2 months ago
063-clusteri...	2 months ago
064-interact...	2 months ago
065-assign...	2 minutes ago
066-data-di...	2 months ago

Task 6.5.15: Use plotly express to create a line plot that shows the values of `silhouette_scores` as a function of `n_clusters`. Be sure to label your x-axis "Number of Clusters", your y-axis "Silhouette Score", and use the title "K-Means Model: Silhouette Score vs Number of Clusters".

```
[43]: # Create a line plot of 'silhouette_scores' vs 'n_clusters'
fig = px.line(
    x=n_clusters, y=silhouette_scores, title="Your Title"
)
fig.update_layout(xaxis_title="Your x_label", yaxis_title="Your y_label")
# Don't delete the code below
fig.write_image("images/6-5-15.png", scale=1, height=500, width=700)
fig.show()
```

Your Title

n_clusters	silhouette_scores
2	0.9
4	0.75
6	0.75
8	0.7
10	0.65
12	0.65

```
[44]: with open("images/6-5-15.png", "rb") as file:
wqet_grader.grade("Project 6 Assessment", "Task 6.5.15", file)
```

Single Python 3 (pykernel) | Idle

Mode: Command Ln 1, Col 1 English (United States) 065-assignment.ipynb

30°C Haze

Search

Applied Data Science Lab | My Path Module | WorldQuant | work/ds-curricu - JupyterLab | Data-Science-Lab/model.py at n | (45) YouTube

Applied Data Science Lab | My Path Module | WorldQuant | work/ds-curricu - JupyterLab | Data-Science-Lab/model.py at n | (45) YouTube

vm.wqu.edu/lab/tree/work/ds-curriculum/060-consumer-finances-in-usa/065-assignment.ipynb

File Edit View Run Kernel Tabs Settings Help

065-assignment.ipynb

Filter files by name

Name	Last Modified
065-assignment.ipynb	3 hours ago
data	2 months ago
images	2 minutes ago
061-explorin...	2 months ago
062-clusteri...	2 months ago
063-clusteri...	2 months ago
064-interact...	2 months ago
065-assign...	in a few seconds
066-data-di...	2 months ago

Task 6.5.16: Build and train a new k-means model named `final_model1`. The number of clusters should be 3.

Note: For reproducibility, make sure you set the random state for your model to 42.

```
[45]: final_model = make_pipeline(
StandardScaler(),
KMeans(n_clusters=3, random_state=42)
)
final_model.fit(X)
```

Pipeline

- StandardScaler
- KMeans

```
[46]: # match_steps, match_hyperparameters, prune_hyperparameters should all be True
wqet_grader.grade("Project 6 Assessment", "Task 6.5.16", final_model)
```

Way to go!
Score: 1

Communicate

Excellent! Let's share our work!

Task 6.5.17: Create a DataFrame `xgb` that contains the mean values of the features in `X` for the 3 clusters in your `final_model1`.

```
[ ]: labels = ...
```

Single Python 3 (pykernel) | Idle

Mode: Command Ln 1, Col 1 English (United States) 065-assignment.ipynb

30°C Haze

Search

Applied Data Science Lab | My Path Module | WorldQuant | work/ds-curricu - JupyterLab | Data-Science-Lab/model.py at n | (45) YouTube

Applied Data Science Lab | My Path Module | WorldQuant | work/ds-curriculum - JupyterLab | Data-Science-Lab/6_communicate | (45) YouTube

vm.wqu.edu/lab/tree/work/ds-curriculum/060-consumer-finances-in-usa/065-assignment.ipynb

File Edit View Run Kernel Tabs Settings Help

065-assignment.ipynb

Filter files by name

Name	Last Modified
.ipynb_checkpoints	3 hours ago
data	2 months ago
images	seconds ago
061-explorin...	2 months ago
062-cluster...	2 months ago
063-cluster...	2 months ago
064-interact...	2 months ago
065-assign...	2 minutes ago
066-data-di...	2 months ago

Communicate

Excellent! Let's share our work!

Task 6.5.17: Create a DataFrame `xgb` that contains the mean values of the features in `X` for the 3 clusters in your `final_model`.

```
[47]: labels = final_model.named_steps["kmeans"].labels_
      xgb = X.groupby(labels).mean()
      xgb
```

	BUS	NHNFIN	NFIN	NETWORTH	ASSET
0	7.367185e+05	1.002199e+06	1.487967e+06	2.076003e+06	2.281249e+06
1	6.874479e+07	8.202115e+07	9.169652e+07	1.134843e+08	1.167529e+08
2	1.216152e+07	1.567619e+07	1.829123e+07	2.310024e+07	2.422602e+07

```
[48]: wqet_grader.grade("Project 6 Assessment", "Task 6.5.17", xgb)
```

✓ You're making this look easy. 🍌
Score: 1

As usual, let's make a visualization with the DataFrame.

Task 6.5.18: Use `plotly` express to create a side-by-side bar chart from `xgb` that shows the mean of the features in `X` for each of the clusters in your `final_model`. Be sure to label the x-axis `"Cluster"`, the y-axis `"Value ($)"`, and use the title `"Small Business Owner Finances by Cluster"`.

```
[49]: # Create side-by-side bar chart of 'xgb'
fig = px.bar(
    xgb,
    barmode="group",
    title="Your Title"
)
fig.update_layout(xaxis_title="Your x_label", yaxis_title="Your y_label")
# Don't delete the code below 🍌
fig.write_image("images/6-5-18.png", scale=1, height=500, width=700)
fig.show()
```

30°C
Haze

Applied Data Science Lab | My Path Module | WorldQuant | work/ds-curriculum - JupyterLab | Data-Science-Lab/6_communicate | (45) YouTube

vm.wqu.edu/lab/tree/work/ds-curriculum/060-consumer-finances-in-usa/065-assignment.ipynb

File Edit View Run Kernel Tabs Settings Help

065-assignment.ipynb

Filter files by name

Name	Last Modified
.ipynb_checkpoints	3 hours ago
data	2 months ago
images	seconds ago
061-explorin...	2 months ago
062-cluster...	2 months ago
063-cluster...	2 months ago
064-interact...	2 months ago
065-assign...	2 minutes ago
066-data-di...	2 months ago

Task 6.5.18: Use `plotly` express to create a side-by-side bar chart from `xgb` that shows the mean of the features in `X` for each of the clusters in your `final_model`. Be sure to label the x-axis `"Cluster"`, the y-axis `"Value ($)"`, and use the title `"Small Business Owner Finances by Cluster"`.

```
[49]: # Create side-by-side bar chart of 'xgb'
fig = px.bar(
    xgb,
    barmode="group",
    title="Your Title"
)
fig.update_layout(xaxis_title="Your x_label", yaxis_title="Your y_label")
# Don't delete the code below 🍌
fig.write_image("images/6-5-18.png", scale=1, height=500, width=700)
fig.show()
```

variable

- BUS
- NHNFIN
- NFIN
- NETWORTH
- ASSET

100M
50M
0

Your y_label

0 1 2

Your x_label

```
[50]: with open("images/6-5-18.png", "rb") as file:
      wqet_grader.grade("Project 6 Assessment", "Task 6.5.18", file)
```

Applied Data Science LabMy Path Module | WorldQuantUwork/ds-curricu - JupyterLabData-Science-Lab/6_communica(45) YouTube

vm.wqu.edu/lab/tree/work/ds-curriculum/060-consumer-finances-in-usa/065-assignment.ipynb

File Edit View Run Kernel Tabs Settings Help

065-assignment.ipynb

Markdown

Python 3 (pykernel)

Filter files by name

Name	Last Modified
.ipynb_chec...	3 hours ago
data	2 months ago
images	seconds ago
061-explorin...	2 months ago
062-cluster...	2 months ago
063-cluster...	2 months ago
064-interact...	2 months ago
065-assign...	2 minutes ago
066-data-di...	2 months ago

Task 6.5.19: Create a PCA transformer, use it to reduce the dimensionality of the data in `X` to 2, and then put the transformed data into a DataFrame named `X_pca`. The columns of `X_pca` should be named `"PC1"` and `"PC2"`.

```
[51]: pca = PCA(n_components=2, random_state=42)
X_t = pca.fit_transform(X)
X_pca = pd.DataFrame(X_t, columns=["PC1", "PC2"])

[52]: wqet_grader.grade("Project 6 Assessment", "Task 6.5.19", X_pca)
```

Very impressive.

Score: 1

Finally, let's make a visualization of our final DataFrame.

Task 6.5.20: Use `plotly express` to create a scatter plot of `X_pca` using `seaborn`. Be sure to color the data points using the labels generated by your `final_model1`. Label the x-axis `"PC1"`, the y-axis `"PC2"`, and use the title `"PCA Representation of Clusters"`.

```
[53]: # Create scatter plot of 'PC2' vs 'PC1'
fig = px.scatter(
    data_frame=X_pca,
    x="PC1",
    y="PC2",
    color=labels.astype(str),
    title="PCA Representation of Clusters"
)
fig.update_layout(xaxis_title="PC1", yaxis_title="PC2")
# Don't delete the code below
fig.write_image("images/6-5-20.png", scale=1, height=500, width=700)
fig.show()
```

Simple 0 1 Python 3 (pykernel) | Idle Mode: Command Ln 1, Col 1 English (United States) 065-assignment.ipynb

30°C Haze Search ENG IN 18:47 18-03-2023

Applied Data Science LabMy Path Module | WorldQuantUwork/ds-curricu - JupyterLabData-Science-Lab/6_communica(45) YouTube

vm.wqu.edu/lab/tree/work/ds-curriculum/060-consumer-finances-in-usa/065-assignment.ipynb

File Edit View Run Kernel Tabs Settings Help

065-assignment.ipynb

Markdown

Python 3 (pykernel)

Filter files by name

Name	Last Modified
.ipynb_chec...	3 hours ago
data	2 months ago
images	seconds ago
061-explorin...	2 months ago
062-cluster...	2 months ago
063-cluster...	2 months ago
064-interact...	2 months ago
065-assign...	2 minutes ago
066-data-di...	2 months ago

PCA Representation of Clusters

```
[54]: with open("images/6-5-20.png", "rb") as file:
wqet_grader.grade("Project 6 Assessment", "Task 6.5.20", file)
```

You got it. Dance party time!

Score: 1

Copyright 2022 WorldQuant University. This content is licensed solely for personal use. Redistribution or publication of this material is strictly prohibited.

Simple 0 1 Python 3 (pykernel) | Idle Mode: Command Ln 1, Col 1 English (United States) 065-assignment.ipynb

30°C Haze Search ENG IN 18:47 18-03-2023