Predicting per capita income percentile group

Content

- Business Use Case
- The NHIES 2016 Dataset Used
- Architectural Decisions
- Process Model Task 1 – Initial Data Exploration
- Process Model Task 2 – Extract, Transform, Load (ETL)
- Process Model Task 3 – Feature Creation/Engineering
- Process Model Task 4 – Model Definition
- Process Model Task 5 – Model Training
- Process Model Task 6 – Model Evaluation
- Process Model Task 7 – Model Deployment
- Conclusions

Business Use Case

- In many countries, public decision makers would want to know more about the factors that mostly correlate with household income.
- Knowing these factors will allow them to put in place programs and projects for poverty alleviation for instance.
- Development institutions are also interested in being able to look at factors influencing income and classify households into income groups to better address their needs.
- In both cases, data science techniques with machine learning algorithms can help achieve the objectives in an efficient and cost-effective manner.

The NHIES 2016 Dataset Used

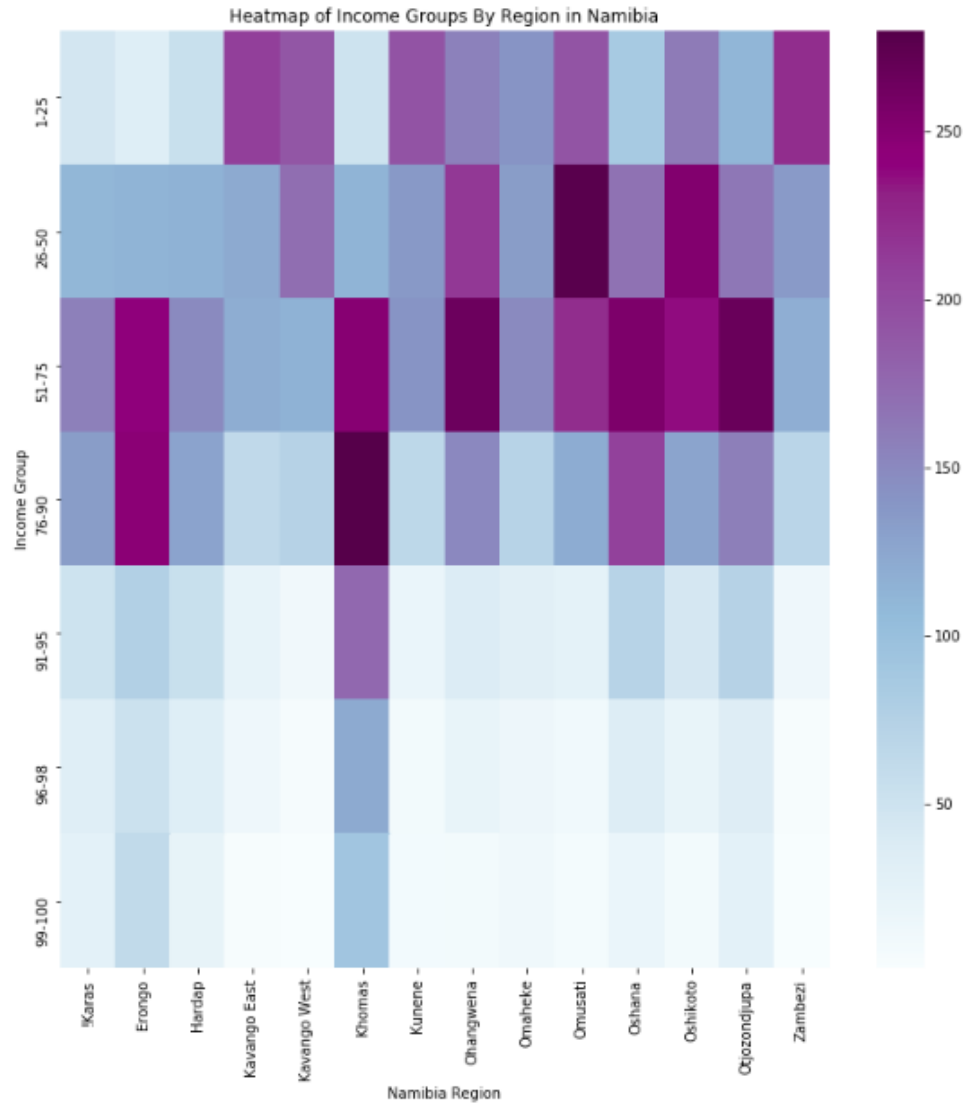
- The Namibia Household Income/Expenditure Survey of 2015-2016 was used. It is publicly available in a anonymized form at https://nsa.org.na/page/central-data-catalogue/
- There are 10090 rows(or observations, samples) and 2485 columns (or variables many with household income related features).
- The key dependent variable –label-- to predict is the adjusted per capita income percentiles (‘apci_groups’).
- In the next slide the Heat Map shows the data with label cross-tabulated against regions. Looking at the map, for instance, one can see that most high income groups are in Khomas region.

NhiesHeatMapByRegion.ipy | Ads_Capstone.feature_selecl | Ads_Capstone.feature_engir

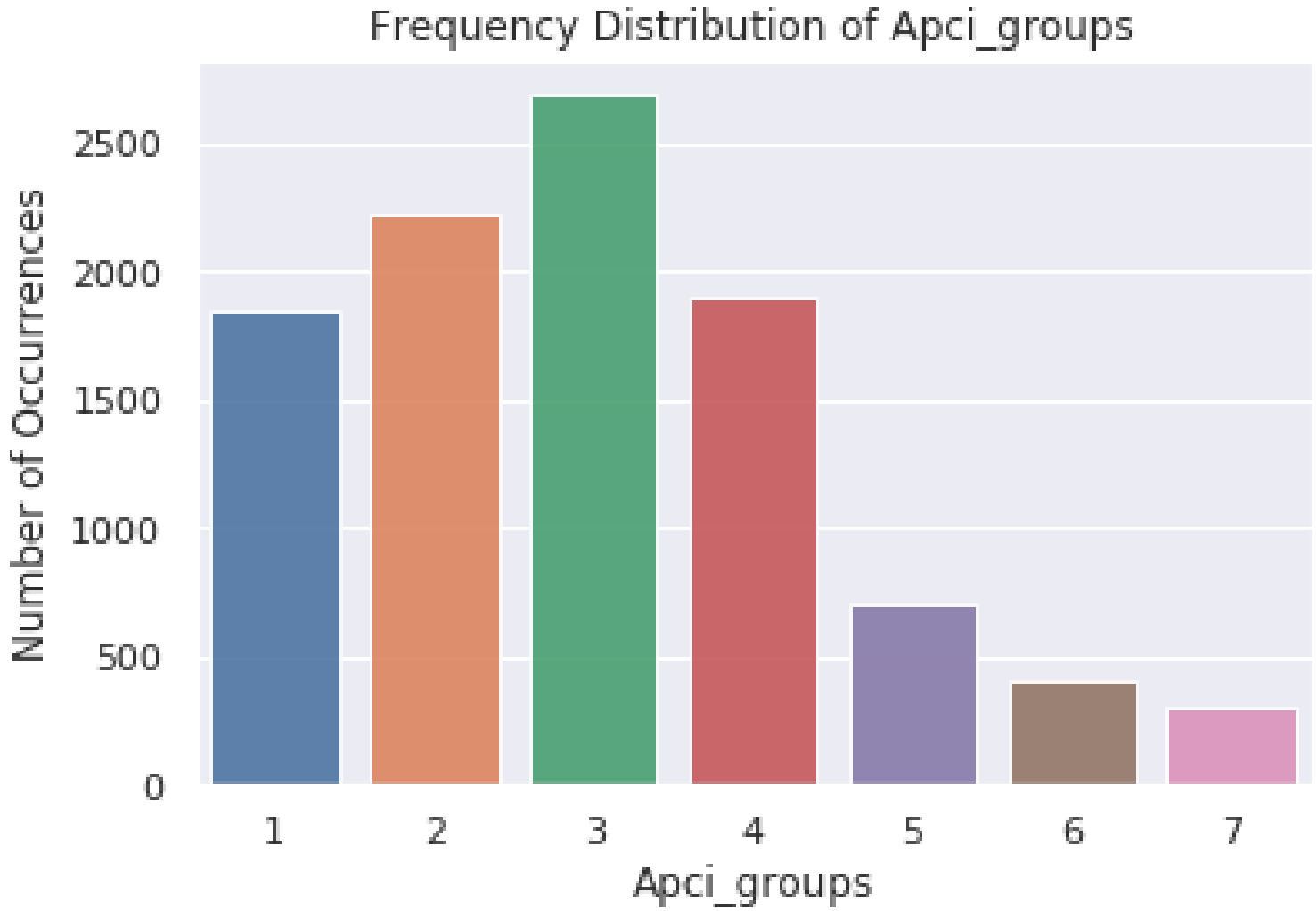
Code

[12]:
```
import seaborn as sns
fig = plt.figure(figsize=(12,12))
r = sns.heatmap(apci_region, cmap='BuPu')
r.set_title("Heatmap of Income Groups By Region in Namibia")
```

[12]: Text(0.5, 1.0, 'Heatmap of Income Groups By Region in Namibia')

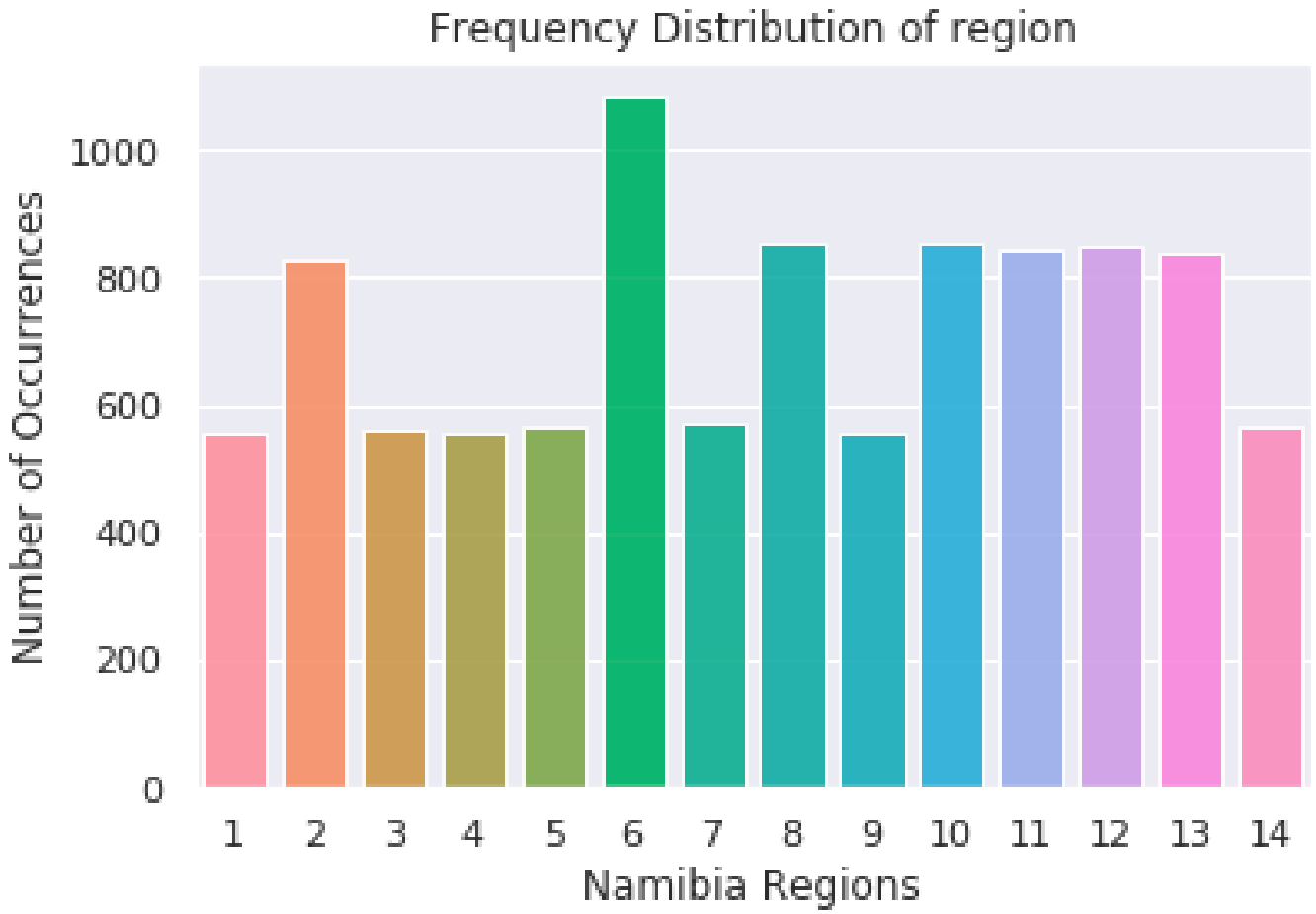


Apci_groups bar chart



1:1-25
2:26-50
3:51-75
4:76-90
5:91-95
6:96-98
7:99-100

Region bar chart



1: !Karas
2: Erongo
3: Hardap
4: Kavango East
5: Kavango West
6: Khomas
7: Kunene
8: Ohangwena
9: Omaheke
10: Omusati
11: Oshana
12: Oshikoto
13: Otjozondjupa
14: Zambezi

Architectural Decisions

- As already stated data set downloaded from NSA Data Portal
  - Data publicly available
  - Anonymized
  - Cleaned
- The process model was implemented on the IBM Watson and IBM Developer Skills Network Labs (CCLabs) on the Cloud.
  - Open source tools used (Jupyter, Python, pandas,matplotlib, scikit-learn, Keras, TensorFlow).
  - There is the possibility to utilize SystemML / Apache Spark to process data in multiple servers – was not used, though. Only part of ETL on Apache Spark.

How was it done?

- By following a Process Model involving the following steps
  - Process Model Task 1 – Initial Data Exploration
  - Process Model Task 2 – Extract, Transform, Load (ETL)
  - Process Model Task 3 – Feature Creation/Engineering
  - Process Model Task 4 – Model Definition
  - Process Model Task 5 – Model Training
  - Process Model Task 6 – Model Evaluation
  - Process Model Task 7 – Model Deployment
- The steps before task 4, were used to select the features of the model
- The associations/correlations between the features themselves and the features and the variable to predict were analyzed thoroughly using robust statistical techniques.

Correlations between quantitative variables

Skills Network Labs

File Edit View Run Kernel Tabs Settings Help

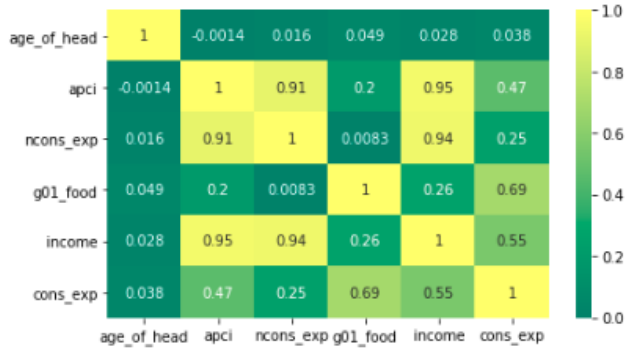
NhiesHeatMapByRegion.ipy | Ads_Capstone.feature_selec | Ads_Capstone.feature_engir

Markdown

[20]:
```
# Heat Map

plt.figure(figsize=(7,4))
sns.heatmap(quantds.corr(),annot=True,cmap='summer')
```

[20]: <matplotlib.axes._subplots.AxesSubplot at 0x7f3350065fd0>



Consumption expenditure (cons_exp) is correlated with income and food expenditur (g01_food).

Income correlates with apci, non consumption expenditure (ncons_exp), and consumption expenditure

apci (adjusted per capita income) of the household correlates strongly with ncons_exp and income

Age of head of household is not correlated with the other variables

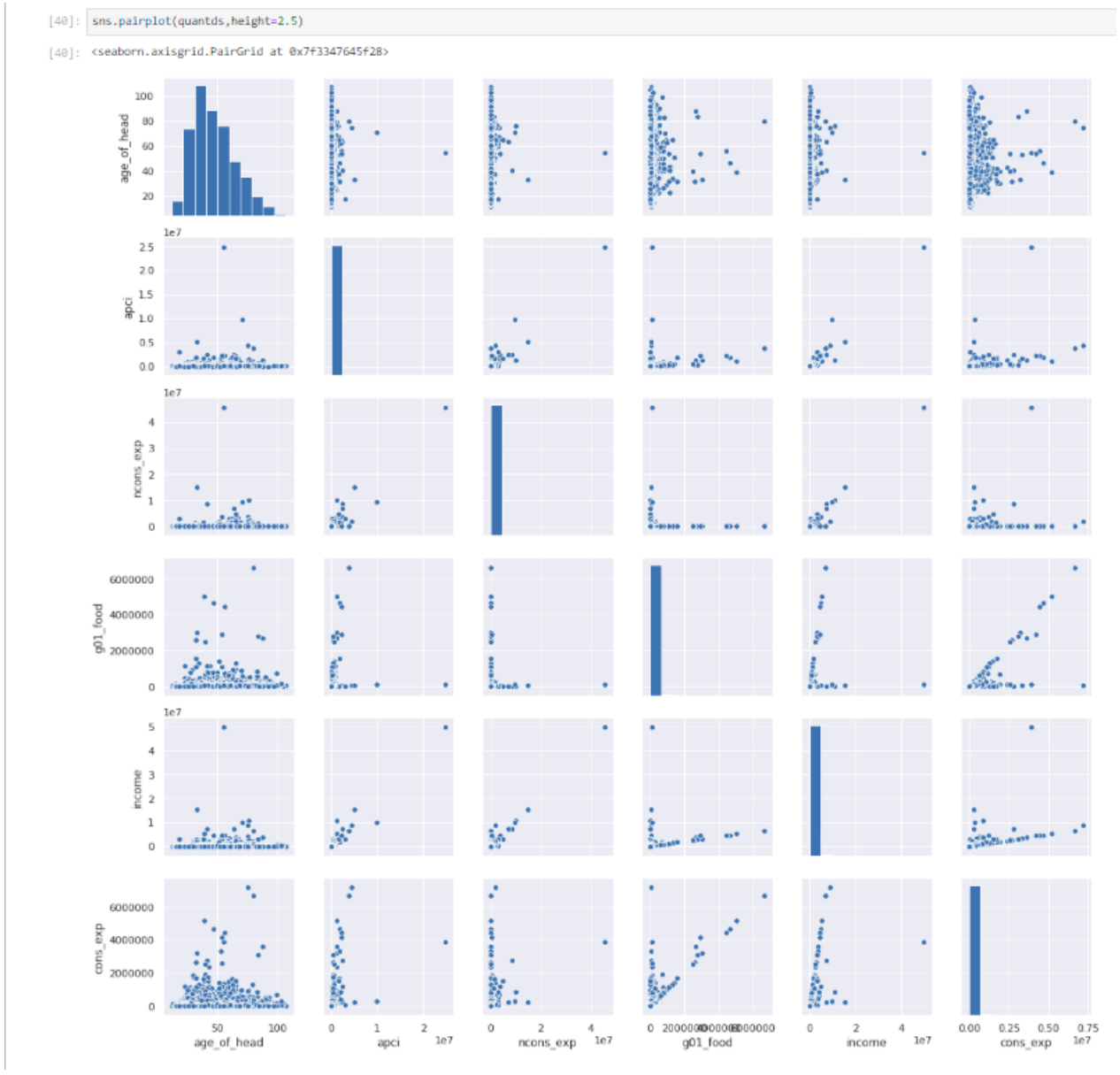
Hence in this study we will limit ourselves to predicting the variable apci_groups (our label or dependent variable --categorical) using

the independent variables (categorical) like age_group, sex_of_head, main_language, main_sourceincome, tenure, urbrur_head, attain

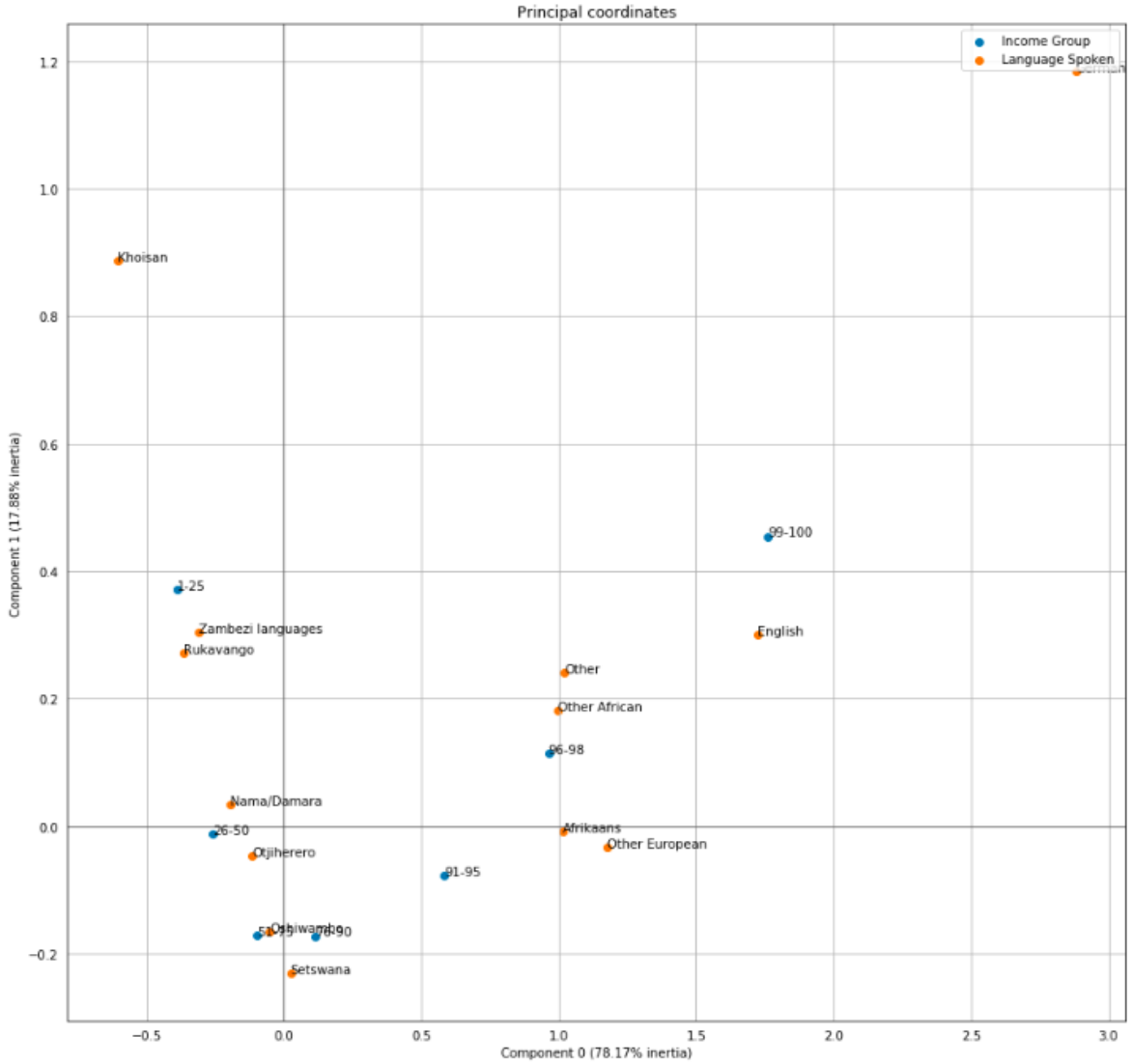
Pair plot between quantitative variables

```
[40]: sns.pairplot(quantds,height=2.5)

[40]: <seaborn.axisgrid.PairGrid at 0x7f3347645f28>
```



Association between categorical variables



How was the Model defined?

- Based on the detective work performed during the tasks 1,2,3, the best potential features selected to predict our label/dependent – ‘apci_groups’ -- variable of interest were: ‘sex_of_head’, ‘income_source’, ‘language’
- Hence in formal terms our equation was:
- apci_groups = f(sex_of_head, income_source, language)
- In practice we will fit this model using Deep Learning technologies (Keras, TensorFlow, Theano) and non Deep Learning technologies such as kNN, GBTClassifier, Multinomial LogisticsRegression, AdaBoostClassifier for comparison purposes.

How well did the model perform?

	Deep Learning	Non Deep Learning				
	Sequential Model	kNN	Logistic Regression	GBTClassifier	AdaBoosterClassifier	RandomForestClassifier
Model Accuracy	0.980	0.922	0.708	0.970	0.701	0.972

Conclusions

- The machine learning models have been used successfully to gain useful insights from a good quality dataset. We predicted with high accuracy the adjusted per capita income percentile group of a household based on the sex of the head of household, the income source of the household, and the language spoken by the head of household.
- The Deep Learning techniques performed better in terms of classification accuracy than their non deep learning counterparts.

Thank You!