Foreign-born vs Natives in Texas

ADTA 5240 Group 1:

Paloma Leonato

Tserendulam Ichinnorov

Eric Droegemeier

Gregory Ehlinger

Brad Reese

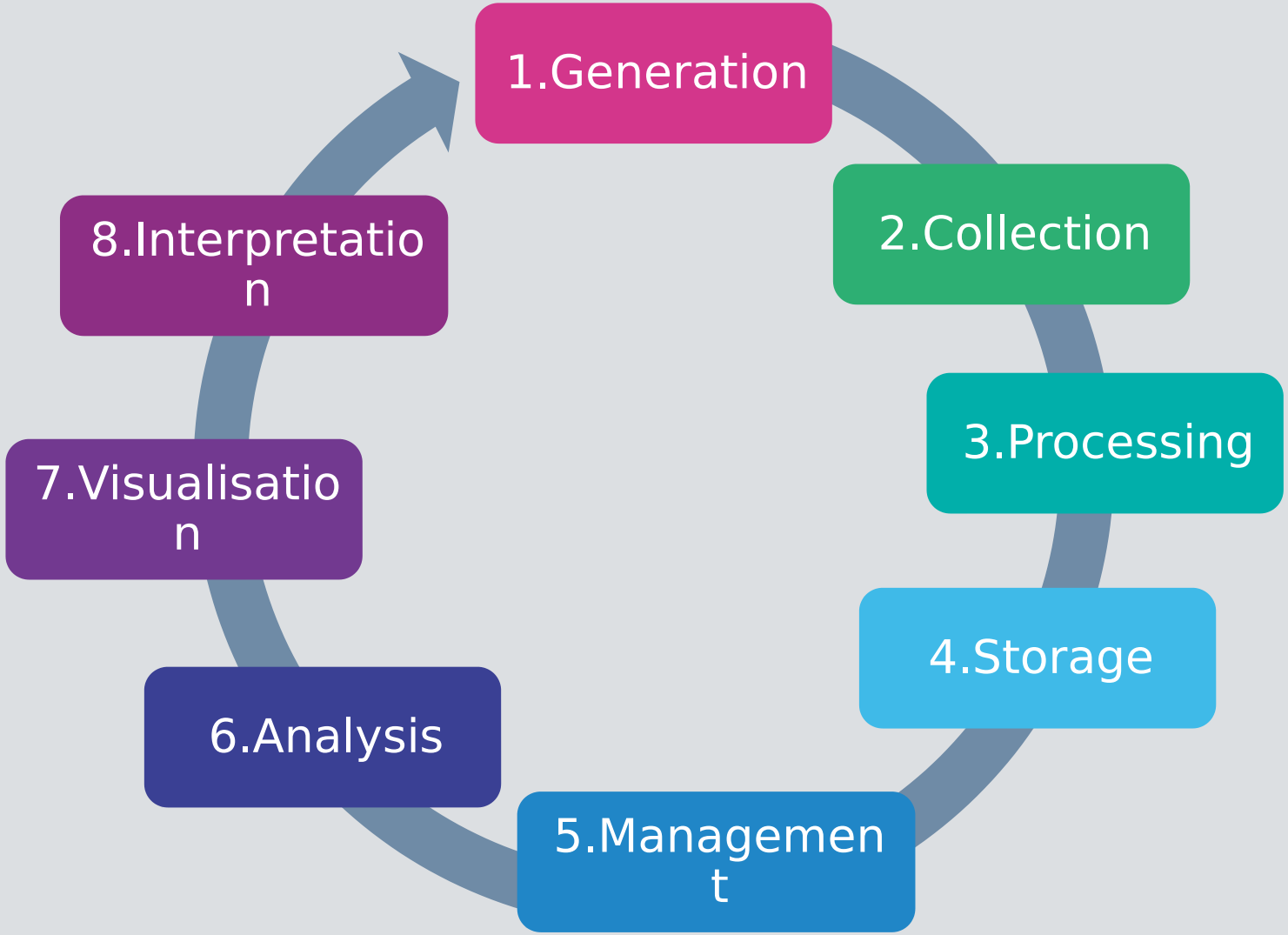
Roles and Responsibilities: All team members share responsibilities

Research Question

What factors differentiate native from foreign born workers of the Texas workforce?

Image from https://www.freepik.com

Analytic life cycle



Generation

Collected: US Census Bureau
https://www.census.gov

Frequency: Every year

Coverage: Sample – 3.5 million households

Coded: By state and US territory

Topics: jobs, occupations, educational attainment, income, veterans, whether people own or rent their homes, and other topics.

Collection

Derived from: US Census Bureau - PUMS (Public Use Microdata sample)

https://www.census.gov/programs-surveys/acs/microdata.html

State: Texas

Year: 2021

Topics: Population: a place of birth, jobs, occupations, income, educational attainment

Records: 261,446 records

Variables: 287

Logo from https://commons.wikimedia.org/wiki/File:US-Census-ACSLogo.svg

Avoiding Bias in the Analysis

Drawbacks of the data

- Encompassing the state of Texas
- American Community Survey is limited to the sample of randomly selected districts in the United States each year
- Outliers can skew data significantly
- Undocumented workers not accounted for

Logos from https://commons.wikimedia.org/

Processing

Cleaning

- Removed data that was not relevant to the study
- From 287 variables to 23
- Limited our scope to people in the working age of 18-65
- Limited our scope to people who work

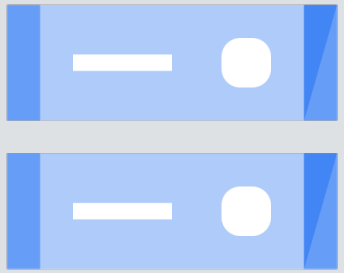
Google Cloud

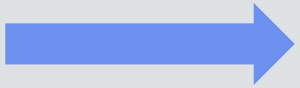
python™

Logos from https://commons.wikimedia.org/

Storage

Google Cloud



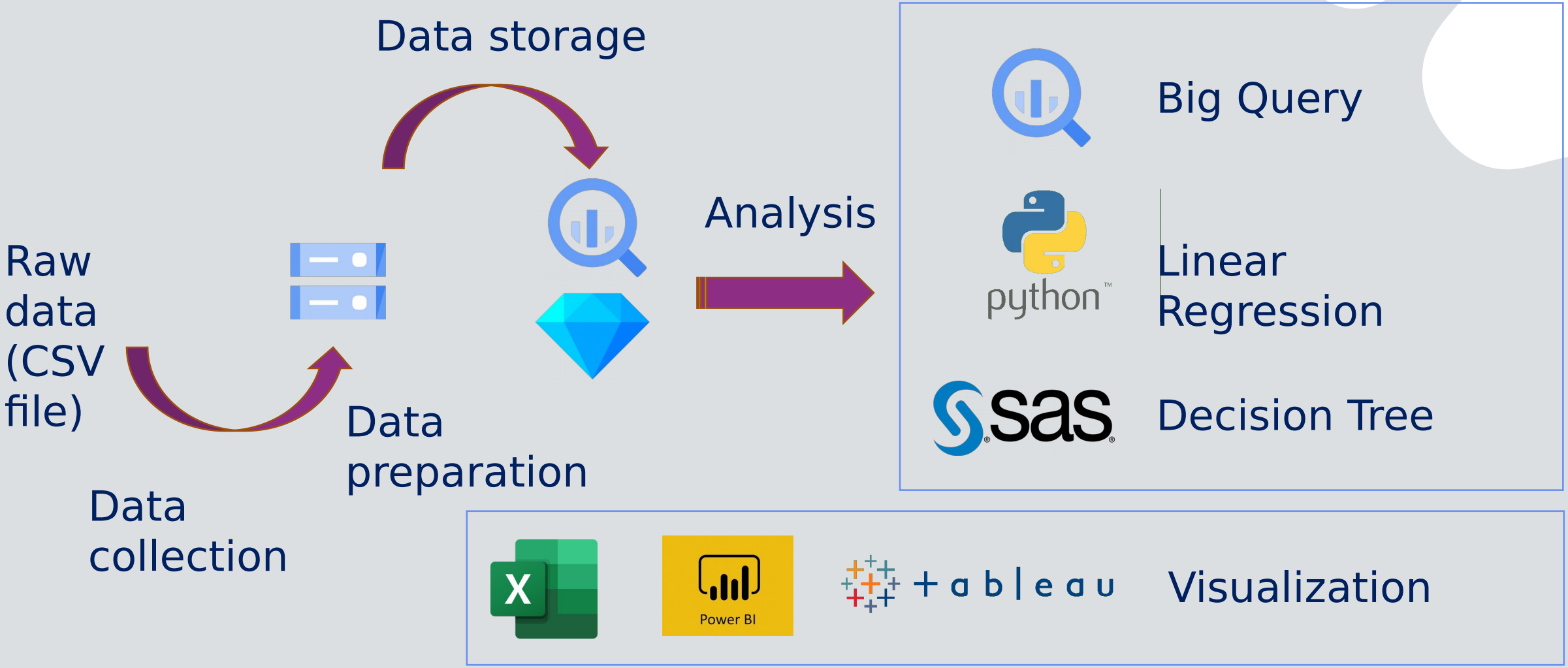


Cloud Storage Bucket

Big Query

Logos from https://cloud.google.com/icons

Management



Logos from https://commons.wikimedia.org/
https://cloud.google.com/icons

Analysis – Big Query – Industry Codes

1

```
SELECT
,CASE
when NAICSP = '115' THEN 'Agriculture'
when NAICSP = '211' THEN 'Oil/Mining'
when NAICSP = '22S' THEN 'Utilities'
when NAICSP = '23' THEN 'Construction'
when NAICSP = '3MS' THEN 'Manufacturing'
when NAICSP = '4231' THEN 'Wholesale'
when NAICSP = '4413' THEN 'Retail'
when NAICSP = '481' THEN 'Transportation'
when NAICSP = '5191ZM' THEN 'Media'
when NAICSP = '5221M' THEN 'Finance'
when NAICSP = '5411' THEN 'Professional Services'
when NAICSP = '611M1' THEN 'Education'
when NAICSP = '623M' THEN 'Medical'
when NAICSP = '6241' THEN 'Social Services'
when NAICSP = '7222' THEN 'Entertainment'
when NAICSP = '811192' THEN 'Services'
when NAICSP = '923' THEN 'Administration'
when NAICSP = '928110P1' THEN 'Military'
when NAICSP = '999920' THEN 'Unemployed' ELSE 'NA'
END as Industry2
Industry, SOCP, WOAB, Industry2, Occupation2, RT,
SOCP, WOAB, Occupation2, ST, NOP, AGEP, POB2, SCHL2
,SUM(PERNP) as Earnings
,SUM(PINCP) as Income
FROM `crucial-decoder-373401.Immigration_01.Texas3`
WHERE RT in('P')
AND AGEP > 18 and AGEP < 66
GROUP BY Industry, SOCP, WOAB, Industry2, Occupation2, RT
,ENGL
,SCHL
,ST
,NOP
,AGEP
,POBP2
,SCHL2
```

2

3

4

5

1. Case statement to convert NAISCP codes into Industry labels – 269 codes mapped
2. Remaining fields added in select statement
3. Income and Earnings are aggregated
4. Filter on "Person" records and working age group
5. Group by all non-aggregated fields

Analysis – Big Query – Occupation Codes

1

```
SELECT
,CASE
when SOCP = '1191XX' THEN 'Manager'
when SOCP = '131011' THEN 'Business'
when SOCP = '1320XX' THEN 'Finance'
when SOCP = '151221' THEN 'Computer Sciences'
when SOCP = '1520XX' THEN 'Computer Sciences'
when SOCP = '171011' THEN 'Engineering'
when SOCP = '191010' THEN 'Science'
when SOCP = '212099' THEN 'Social Work'
when SOCP = '2310XX' THEN 'Legal'
when SOCP = '2590XX' THEN 'Education'
when SOCP = '271010' THEN 'Entertainment'
when SOCP = '299000' THEN 'Medical'
when SOCP = '311121' THEN 'Health'
when SOCP = '33909X' THEN 'Fire/Police'
when SOCP = '351011' THEN 'Restaurants'
when SOCP = '37301X' THEN 'Maintenance'
when SOCP = '391000' THEN 'Personal Care'
when SOCP = '419099' THEN 'Sales'
when SOCP = '431011' THEN 'Admin'
when SOCP = '454020' THEN 'Game and Fish'
when SOCP = '471011' THEN 'Construction'
when SOCP = '4750XX' THEN 'Oil/Mining'
when SOCP = '491011' THEN 'Trades'
when SOCP = '5371XX' THEN 'Transportation'
when SOCP = '999920' THEN 'Military' ELSE 'NA'
END as SOCP
```

2

```
Industry, SOCP, WOAB, Industry2, Occupation2, RT,
SOCP, WOAB, Occupation2, ST, NOP, AGEP, POB2, SCHL2
```

3

```
,SUM(PERNP) as Earnings
,SUM(PINCP) as Income
```

```
FROM `crucial-decoder-379401.Immigration_01.Texas3`
```

4

```
WHERE RT in('P')
AND AGEP > 18 and AGEP < 66
```

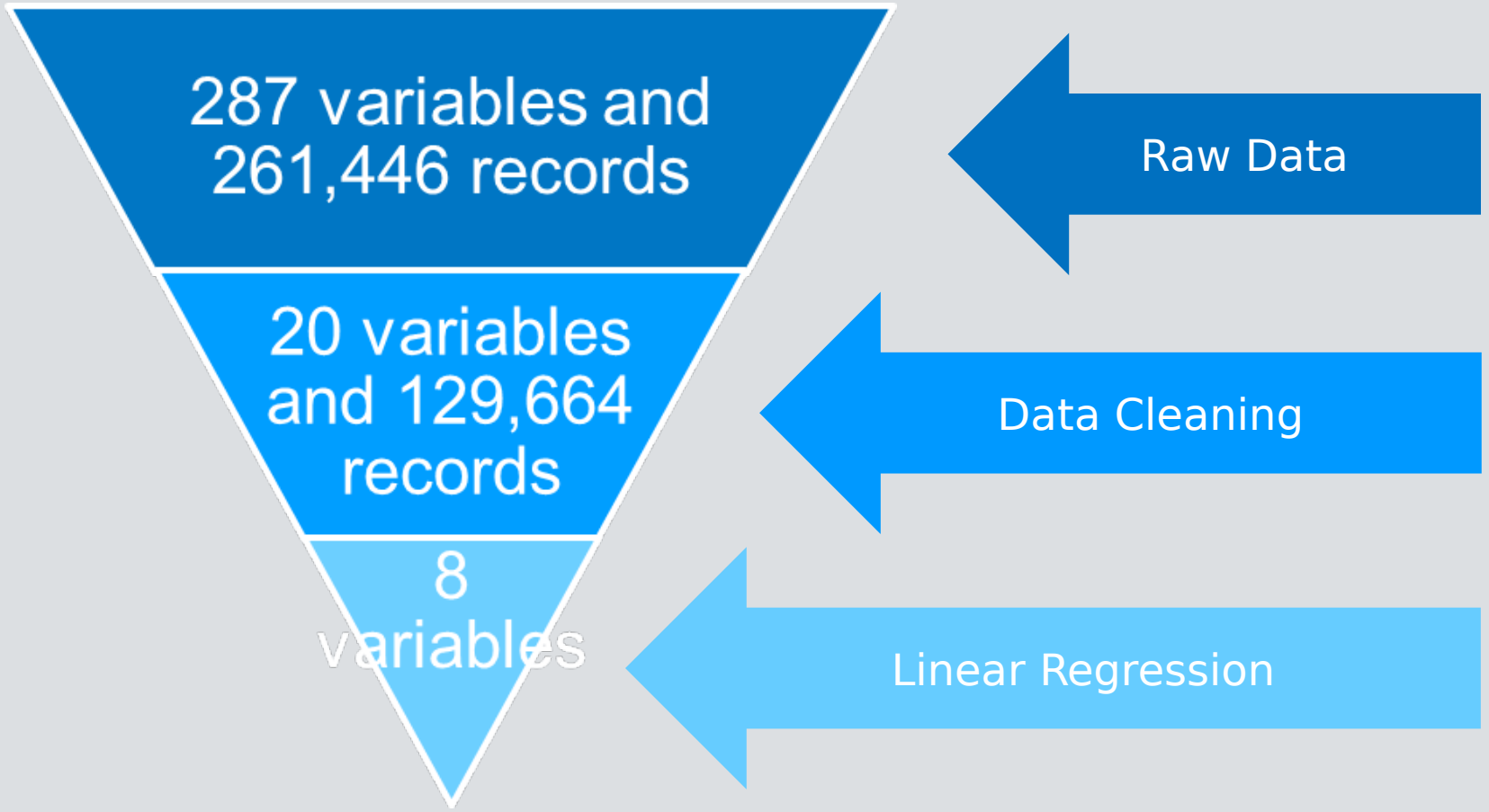
```
GROUP BY Industry, SOCP, WOAB, Industry2, Occupation2, RT
```

5

```
,ENGL
,SCHL
,ST
,NOP
,AGEP
,POBP2
,SCHL2
```

1. Case statement to convert Occupation codes into Industry labels – 529 codes manned
2. Remaining fields added in select statement
3. Income and Earnings are aggregated
4. Filter on "Person" records and working age group
5. Group by all non-aggregated fields

Analysis – Linear regression



Dependent variable =
- **Salary**

Independent numerical variables::
- **Age**
- **hrs/wk**

Dummy variables:
- **Sex**
- **Education**
- **Area of birth**
- **Type of work**

Logos from https://www.flaticon.com

Analysis – Linear regression

Variables used in ML Lineal Regression:

```
dflr = df5[{'AGEP','WKHP','US','AEU','MALE','GRADU','BACHE','SALARYW','OWNERW','WAGP'}]
```

	WKHP	BACHE	AEU	GRADU	AGEP	US	WAGP	OWNERW	MALE	SALARYW
0	25.0	0	0	0	21	1	7000.0	0	1	1
9	30.0	0	0	0	18	1	13000.0	0	1	1
13	6.0	0	0	0	20	1	500.0	0	1	1
14	4.0	0	0	0	18	1	200.0	0	0	1
15	20.0	0	0	0	20	1	1600.0	0	0	1
19	40.0	0	0	0	29	1	3800.0	0	1	1
20	40.0	0	0	0	64	1	18000.0	0	1	1
21	40.0	0	0	0	22	1	22000.0	0	1	1

```
df5['MALE'] = 0

df5.loc[(df5['SEX'] == 1) , 'MALE'] = 1
```

```
df5['BACHE'] = 0
df5['GRADU'] = 0

df5.loc[(df5['SCHL'] == 1) | (df5['SCHL'] == 2) , 'BACHE'] = 1
df5.loc[df5['SCHL'] > 21  , 'GRADU'] = 1
```

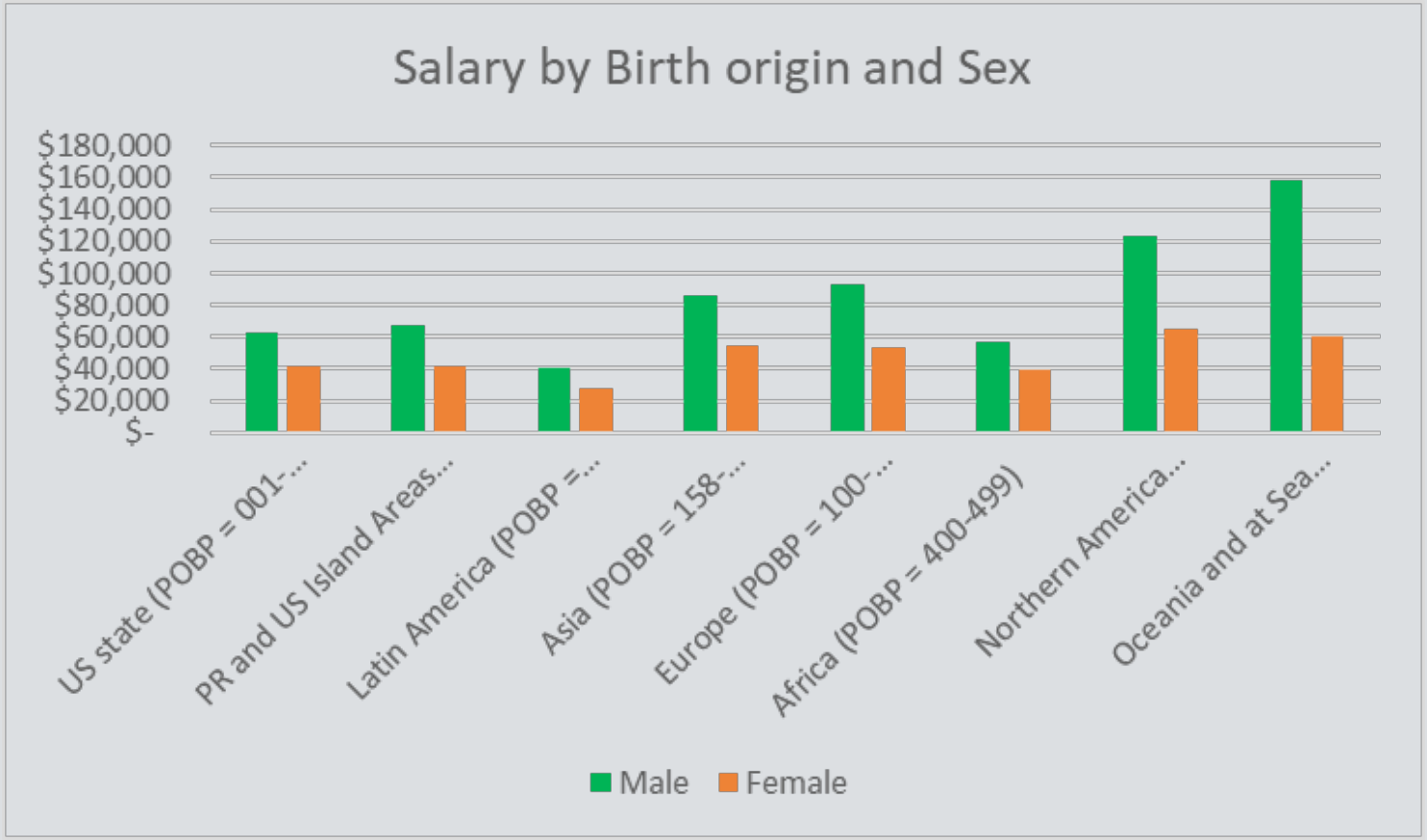
```
df5['SALARYW'] = 0
df5['OWNERW'] = 0

df5.loc[(df5['COW'] <= 5) , 'SALARYW'] = 1
df5.loc[df5['COW'] == 7  , 'OWNERW'] = 1
```

Dummy Variables

Logos from https://www.flaticon.com

Analysis – Linear regression

Lineal Regression equation:



Male's salaries higher than Female in all regions of the world.

Logos from https://www.flaticon.com

Analysis – Linear regression

Lineal Regression equation:

SALARY =

-93,421.23

-14,080*BACHE → Bachelor degree

+11,802*US → Born in the USA

+46,673*SALARYW → Salary worker

+48,783*GRADU → Graduated degree

+1302*WKHP → Number of work hrs/wk

+597*AGEP → Age

+63,193*OWNERW → Own a business

+19,626*AEU → Asian or European

+15,483*MALE → Male

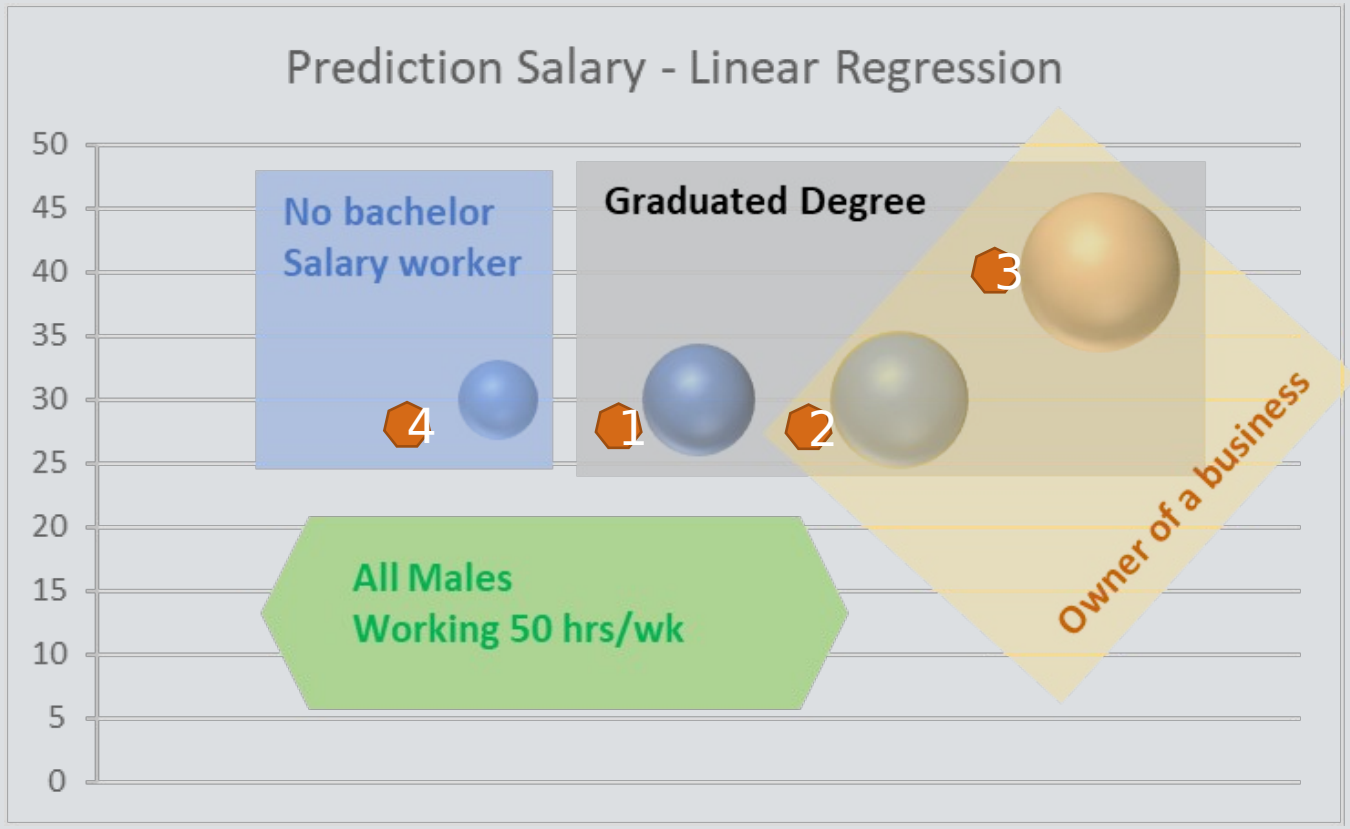
Male salaries higher than Female by $15,000
Business Owners higher contribution to salary
Graduate degrees boost your salary
Foreign born workers higher contribution to salary

R-squared: 0.23681

Logos from https://www.flaticon.com

Analysis – Linear regression

Predictions:



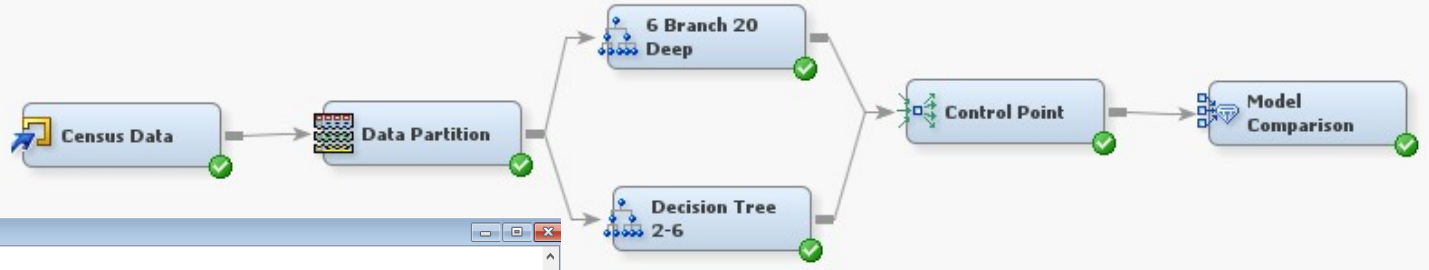
CASE		SALARY
1	$	112,369
2	$	160,079
3	$	163,761
4	$	69,560

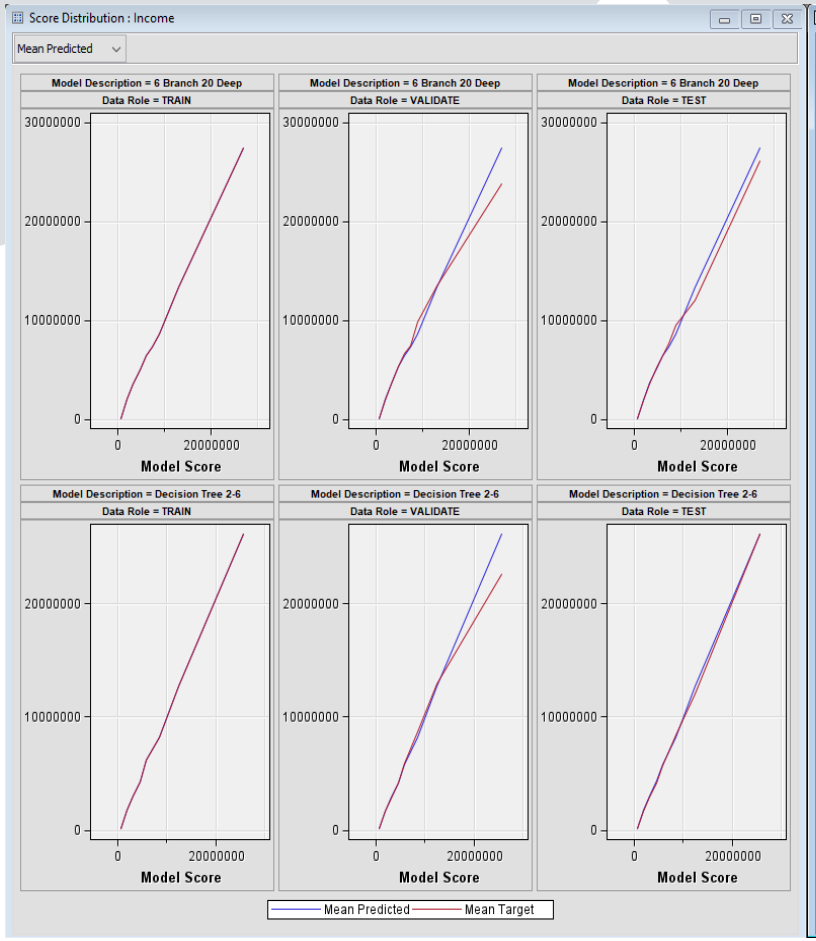
Domestic

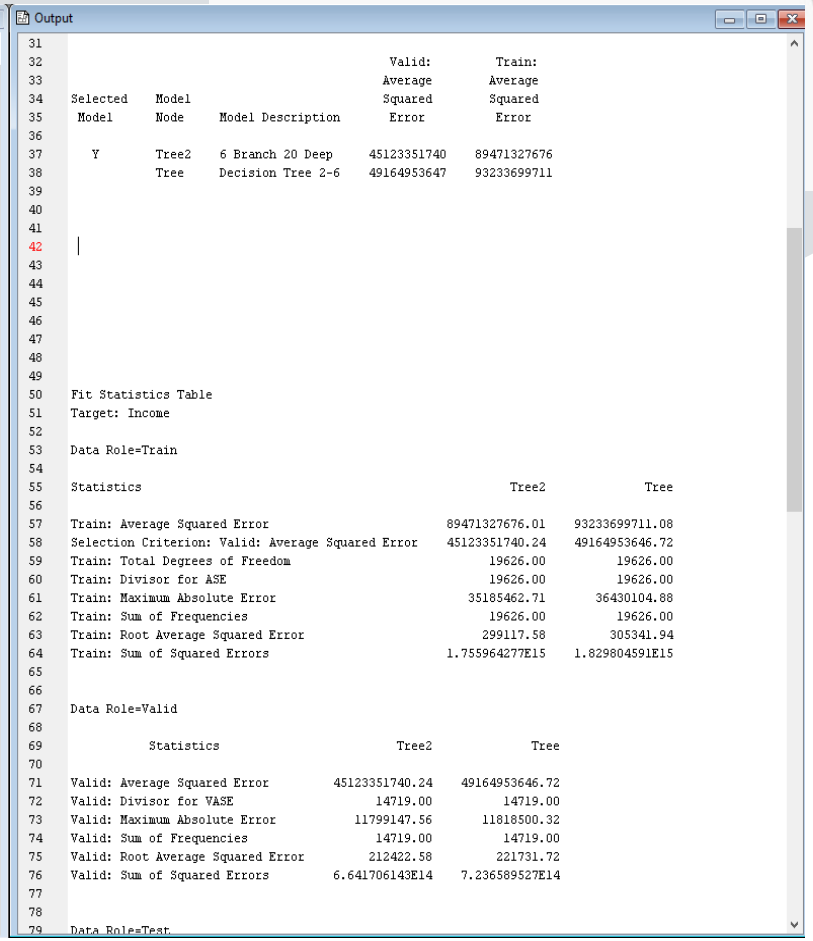
Foreigner

Logos from https://www.flaticon.com

Analysis – Decision Tree

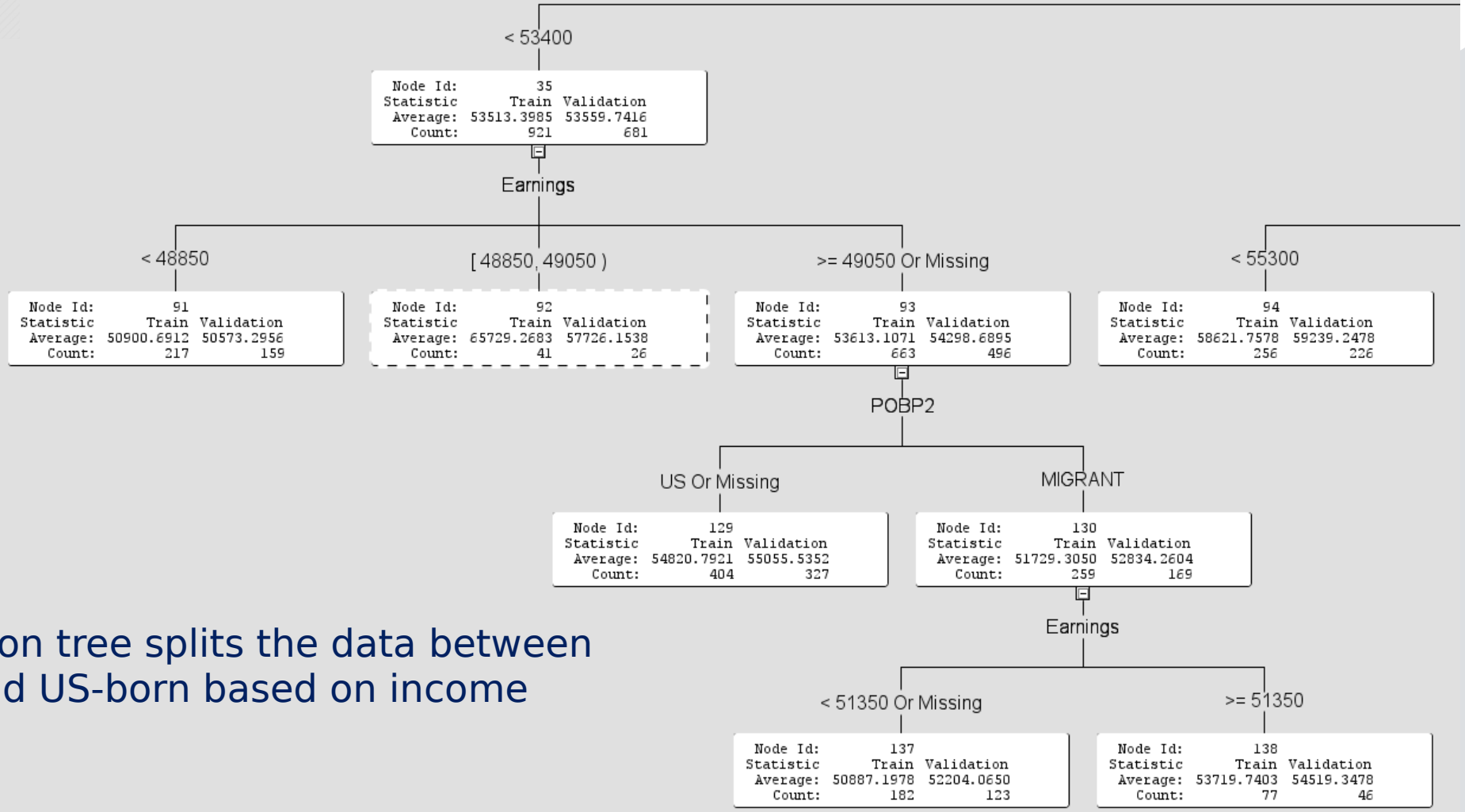






A Decision Tree was run with SAS enterprise miner
Two branches – six depth
Six branches, 20 depth
Model comparison shows the 6-20 tree was the
better fit - stronger Average Square Error

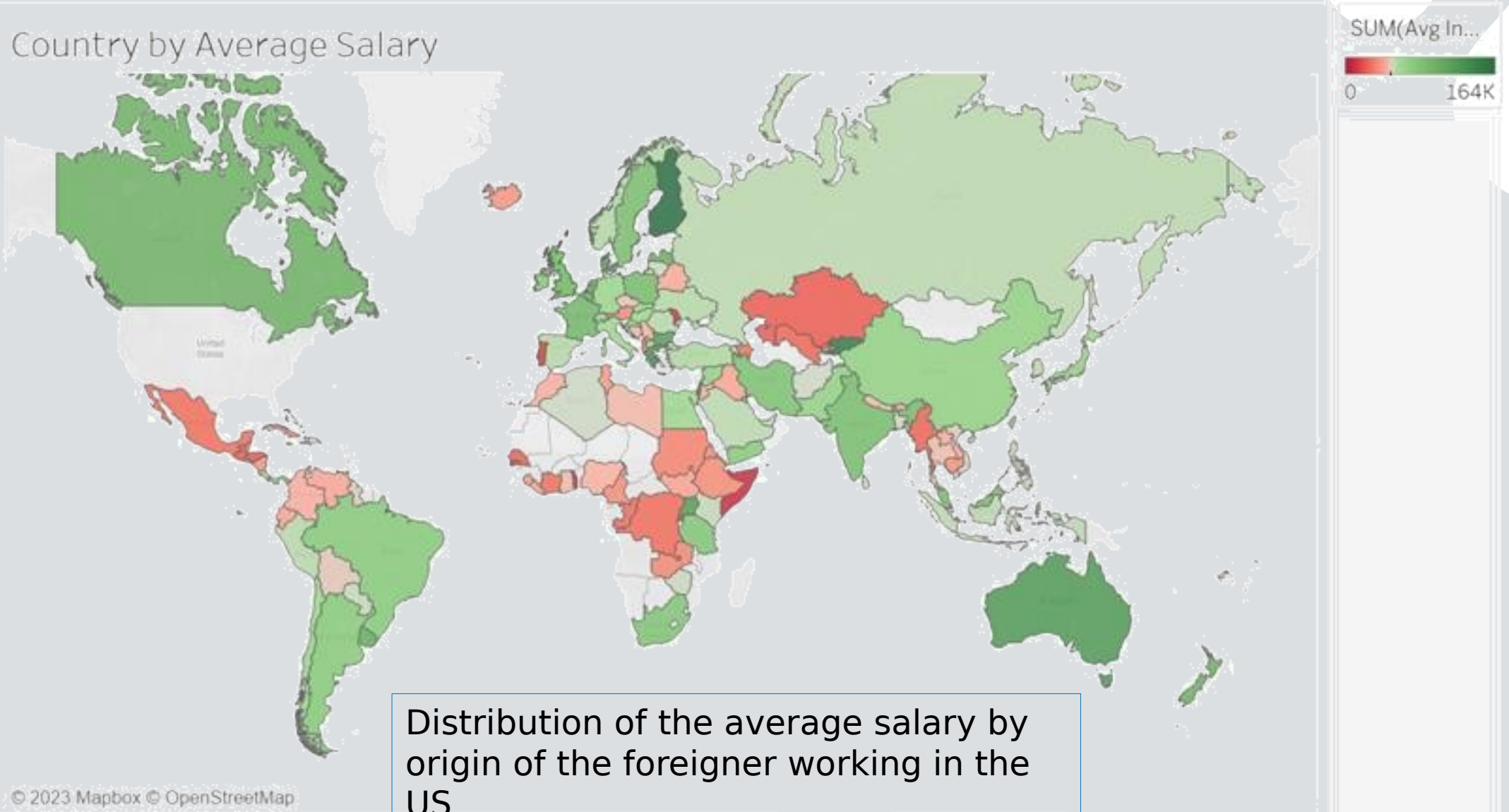
Analysis – Decision Tree



This decision tree splits the data between Migrant and US-born based on income levels.

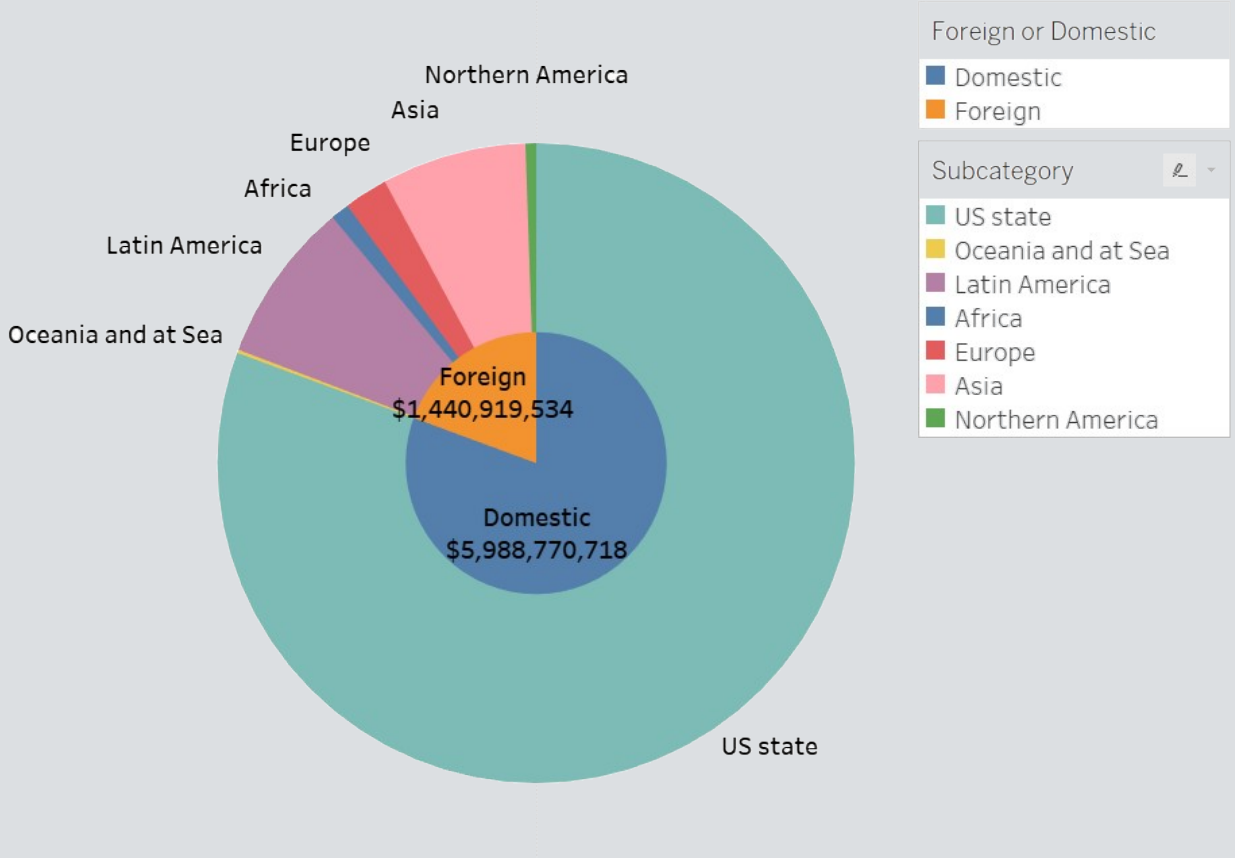
Logos from https://www.flaticon.com

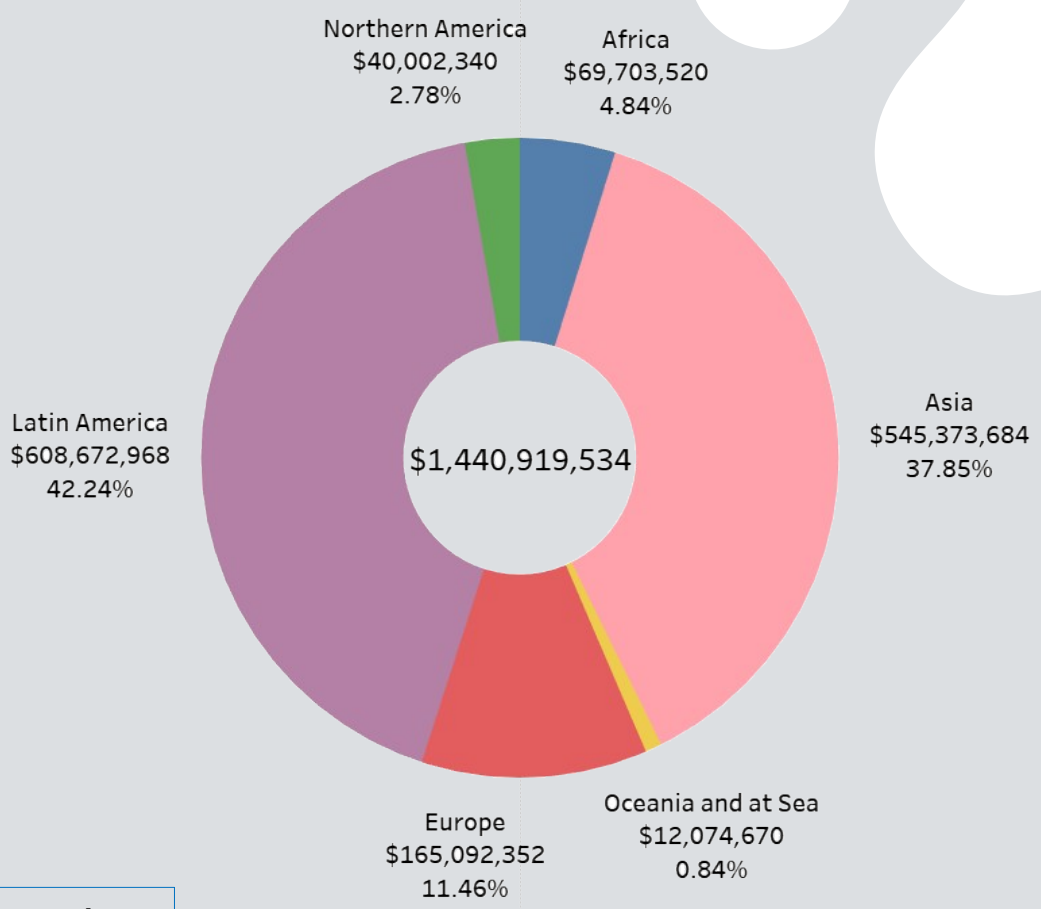
Visualization



Distribution of the average salary by origin of the foreigner working in the US

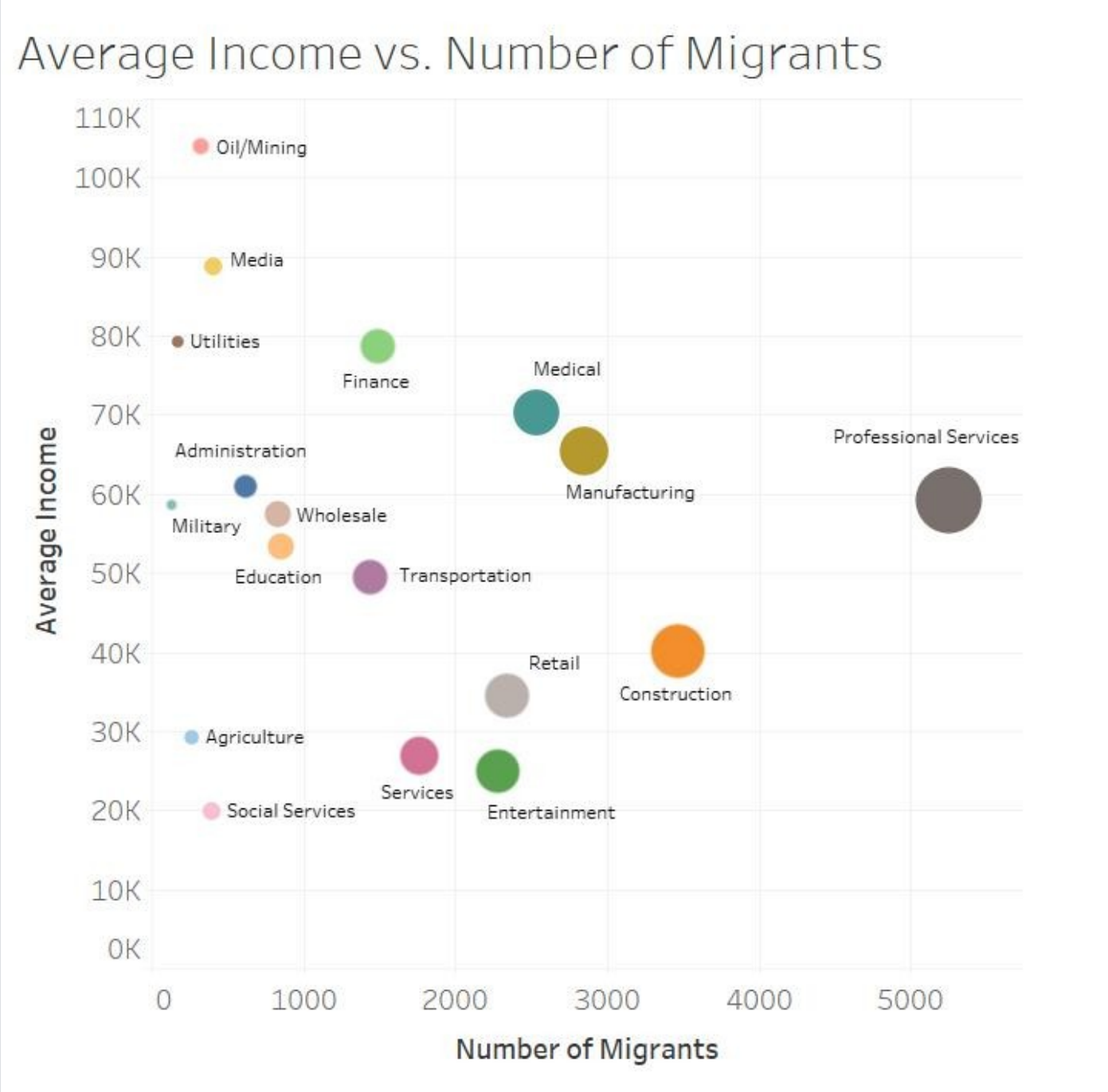
Visualization





The majority of the income in the US is generated by the US born population. Latin America and Asia are the largest contributor to the foreign income.

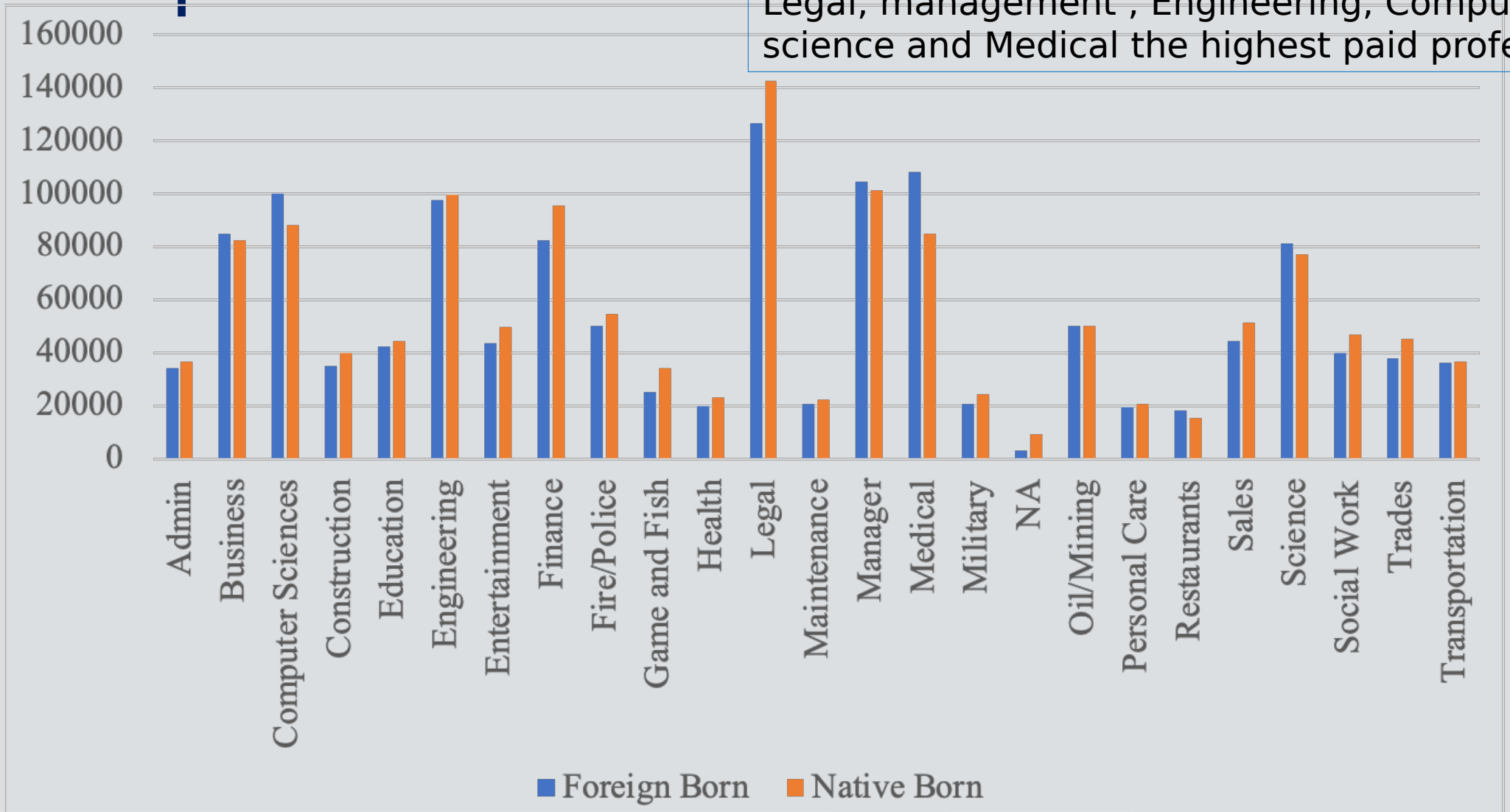
Visualization



Professional Services is one of the most common industry for foreign workers and the average salary in the $60Ks.

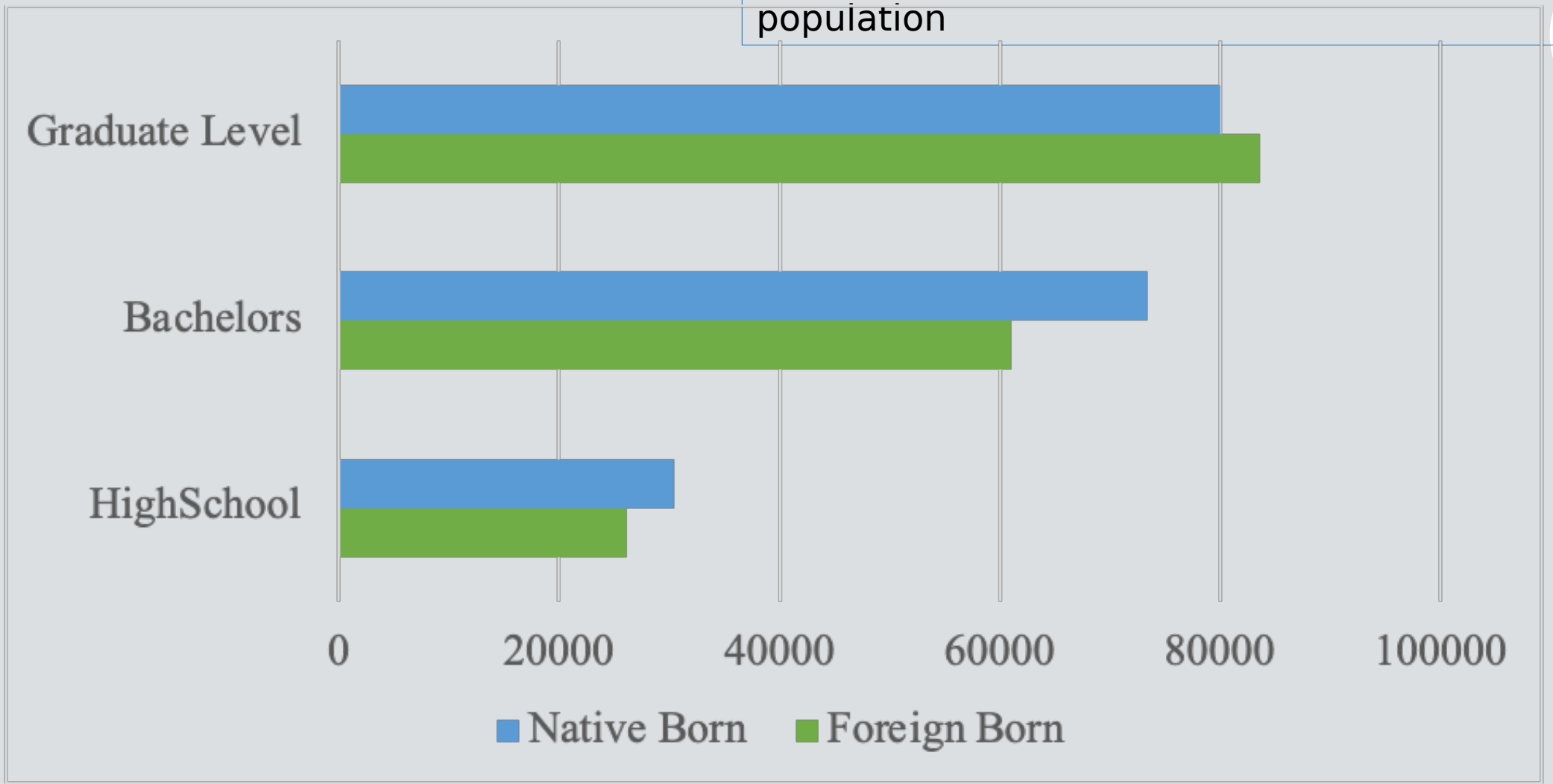
Visualization – Average income by occupation

Legal, management , Engineering, Computer science and Medical the highest paid professions.



Visualization – Average income by education

Foreigners more qualify than US Born population



Interpretation

- Latin America and Asia are the largest foreign regions
- Oil/Mining, Construction, Profession Services industries:
  - Top 2: Northern America and Oceania and at Sea by average income
  - Latin America for the largest group but the lowest average income
- Foreigners qualify more than US Born workers

Interpretation – Average income

By Industry

Foreign born higher than natives:

- Media ($16000 higher)
- Military
- Oil/Mining
- Medical
- Education
- Entertainment

Natives higher than foreign-born:

- Other industries
  - Agriculture ($20000 higher)
  - Wholesale

By Occupation

Foreign born higher than natives:

- Medical ($23000 higher)
- Computer Science
- Science
- Manager
- Restaurant
- Business

Natives higher than foreign-born:

- Other occupations
  - Legal ($15000 higher)
  - Finance
  - Game and Fish, etc.

you

Citations

United States Census Bureau. (n.d.). *Index of /programs-surveys/acs/data/pums*. Index of /programs-surveys/ACS/Data/Pums. Retrieved April 3, 2023, from https://www2.census.gov/programs-surveys/acs/data/pums/

Directions for Supervised Machine Learning Linear and Logistic Regression.pdf by Dr Floyd

Analysis – Big Query

```
SELECT
,CASE
when ENG = 1 THEN 'Very well'
when ENG = 2 THEN 'Well'
when ENG = 3 THEN 'Not well'
when ENG = 4 THEN 'Not at all' ELSE 'NA'
END AS ENGL
```

```
Industry, SOCP, WOAB, Industry2, Occupation2, RT,
SOCP, WOAB, Occupation2, ST, NOP, AGEP, POB2, SCHL2
,SUM(PERNP) as Earnings
,SUM(PINCP) as Income
```

```
FROM `crucial-decoder-373401.Immigration_01.Texas3`
WHERE RT in('P')
AND AGEP > 18 and AGEP < 66
GROUP BY Industry, SOCP, WOAB, Industry2, Occupation2, RT
,ENGL
,SCHL
,ST
,NOP
,AGEP
,POBP2
,SCHL2
```

Analysis – Big Query

```
SELECT
,CASE
when ENG = 1 THEN 'Very well'
when ENG = 2 THEN 'Well'
when ENG = 3 THEN 'Not well'
when ENG = 4 THEN 'Not at all' ELSE 'NA'
END AS ENGL
```

```
Industry, SOCP, WOAB, Industry2, Occupation2, RT,
SOCP, WOAB, Occupation2, ST, NOP, AGEP, POB2, SCHL2
,SUM(PERNP) as Earnings
,SUM(PINCP) as Income
```

```
FROM `crucial-decoder-379401.Immigration_01.Texas3`
WHERE RT in('P')
AND AGEP > 18 and AGEP < 66
GROUP BY Industry, SOCP, WOAB, Industry2, Occupation2, RT
,ENGL
,SCHL
,ST
,NOP
,AGEP
,POBP2
,SCHL2
```

Analysis – Big Query

```
SELECT
,CASE
when AGEP <19 then 'minor'
when AGEP >65 then 'senior'
ELSE 'working age'
END AS AGEP
 Industry, SOCP, WOAB, Industry2, Occupation2, RT,
SOCP, WOAB, Occupation2, ST, NOP, AGEP, POB2, SCHL2
,SUM(PERNP) as Earnings
,SUM(PINCP) as Income
FROM `crucial-decoder-373401.Immigration_01.Texas3`
WHERE RT in('P')
AND AGEP > 18 and AGEP < 66
GROUP BY Industry, SOCP, WOAB, Industry2, Occupation2, RT
,ENGL
,SCHL
,ST
,NOP
,AGEP
,POBP2
,SCHL2
```