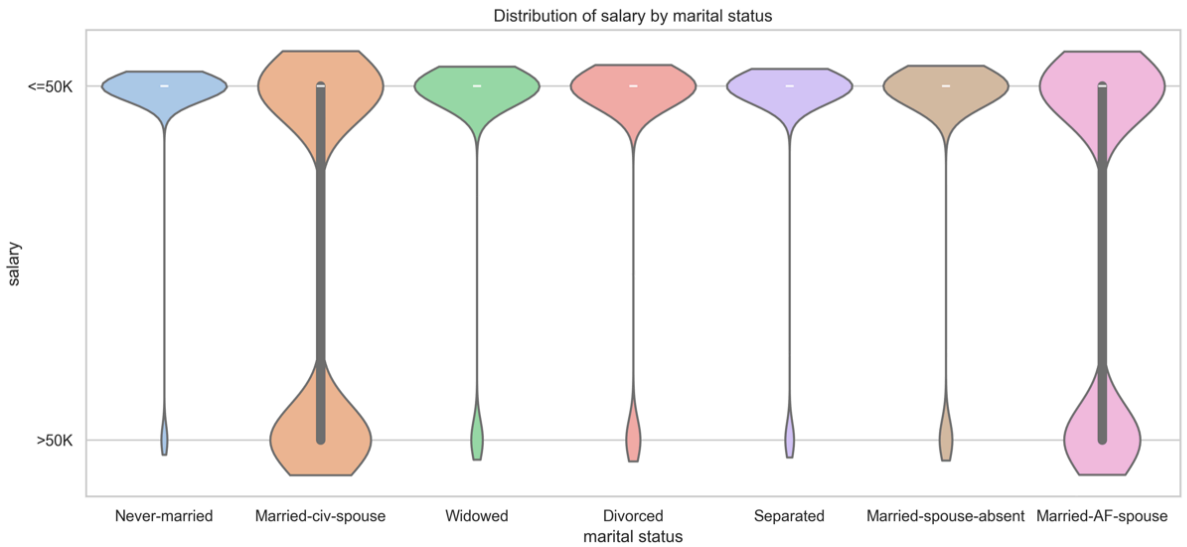
Raport Project1

Niniejszy raport przedstawia analizę zestawu danych Adult Census Income z Kaggle, który zawiera informacje ze spisu powszechnego USA z 2019 r. i był wykorzystany do przewidywania, czy dana osoba zarabia więcej niż 50 000 USD w oparciu o różne cechy społeczno-demograficzne. Zawiera około 49 000 rekordów i 15 atrybutów: wiek, wykształcenie, stan cywilny, zawód, rasę, płeć, narodowość, godziny przepracowane w tygodniu oraz zyski i straty kapitałowe, a także zmienną docelową.

```
rows, columns: (48842, 15)
Analysed numerical data:
                     mean    median           std    min      max  5th_percentile  95th_percentile  missing_values
age                38.643585    37.0     13.710510     17       90           19.00            63.00               0
fnlwgt         189664.134597  178144.5  105604.025423  12285  1490400        39615.40        379481.65               0
education-num      10.078089     10.0      2.570973      1       16            5.00            14.00               0
capital-gain     1079.067626      0.0   7452.019058      0    99999            0.00          5013.00               0
capital-loss       87.502314      0.0    403.004552      0     4356            0.00             0.00               0
hours-per-week     40.422382     40.0     12.391444      1       99           17.05            60.00               0
Analyzed categorical data:
              unique_klasses  missing_values  0_proportion  2_proportion  ...  9_proportion  15_proportion  16_proportion  42_proportion
workclass                  9               0           NaN           NaN  ...           NaN            NaN            NaN            NaN
education                 16               0           NaN           NaN  ...           NaN            NaN            NaN            NaN
marital-status             7               0           NaN           NaN  ...           NaN            NaN            NaN            NaN
occupation                15               0           NaN           NaN  ...           NaN            NaN            NaN            NaN
relationship               6               0           NaN           NaN  ...           NaN            NaN            NaN            NaN
race                       5               0           NaN           NaN  ...           NaN            NaN            NaN            NaN
gender                     2               0           NaN           NaN  ...           NaN            NaN            NaN            NaN
native-country            42               0           NaN           NaN  ...           NaN            NaN            NaN            NaN
salary                     2               0           NaN           NaN  ...           NaN            NaN            NaN            NaN
```

W pierwszej części projektu postanowiłam powtórzyć to doświadczenie i znaleźć zależności między wynagrodzeniami ludności a cechami społeczno-demograficznymi, próbując przewidzieć, czy dana osoba zarabia więcej niż 50 000 USD rocznie. Poniżej przedstawię wykresy pokazujące najważniejsze zależności.

Rozkład wynagrodzeń względem stanu cywilnego




- Osoby pozostające w związku małżeńskim zarabiają ponad 50 tys.

- Osoby, które nigdy nie były w związku małżeńskim, prawie zawsze zarabiają mniej niż 50 tys. zł.

- Osoby owdowiałe, rozwiedzione i w separacji również częściej znajdują się w kategorii <=50 tys. zł.

- Może to sugerować, że stan cywilny wpływa na dochód: małżeństwo, zwłaszcza małżeństwo z konkubentem, koreluje z wyższymi dochodami.

Rozkład wynagrodzeń względem racy

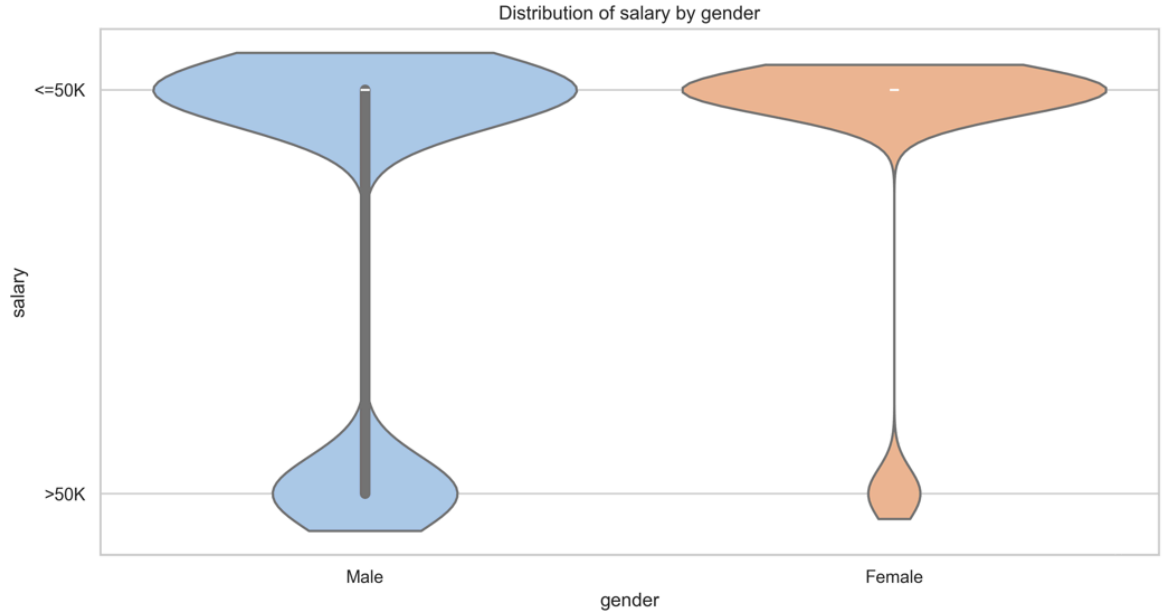

- Z wykresu widać, że większość ludzi we wszystkich kategoriach zarabiają mniej niż 50 000$ w ciągu roku

- Odsetek osób z wynagrodzeniem większym od 50 000$ zupełnie niski wśród amerykańskich Indian i Eskimosów oraz wśród czarnoskórych.

- Biali i Azjaci/Wyspiarze Pacyfiku mają większy odsetek osób z zarobkami powyżej 50 000$, ale nadal stanowią mniejszość.

- Może to wskazywać na różnice rasowe w dochodach, z największą koncentracją wysokich dochodów wśród białych i Azjatów/Wyspiarzy Pacyfiku.

Rozkład wynagrodzeń względem płci



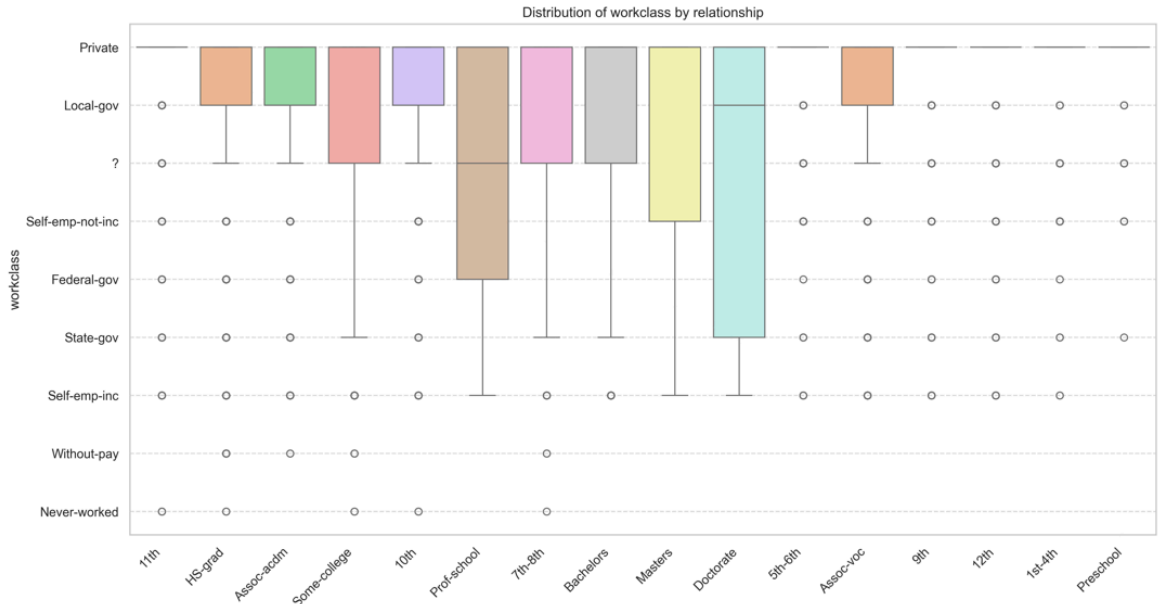

- Mężczyźni dominują w grupie z dochodami (> 50k), podczas gdy kobiety znajdują się w grupie z mniejszymi dochodami (< 50k).

- Wskazuje to na różnicę w dochodach kobiet i mężczyzn, gdzie mężczyźni mają większe szanse na wysokie zarobki.

Na podstawie moich badań i analizy danych mogę powiedzieć, że większość populacji zarabia mniej niż 50 tysięcy rocznie, niezależnie od rasy, stanu cywilnego czy płci.Jednak osoby rasy białej i pochodzącej z Azji i Pacyfiku, osoby zamężne i mężczyźni mają większe szanse na otrzymanie ponad 50 000$

W drugiej części projektu analizowałam dane i próbowałem znaleźć niestandardowe zależności między zestawami.

Rozkład klasy robotniczej według relacji

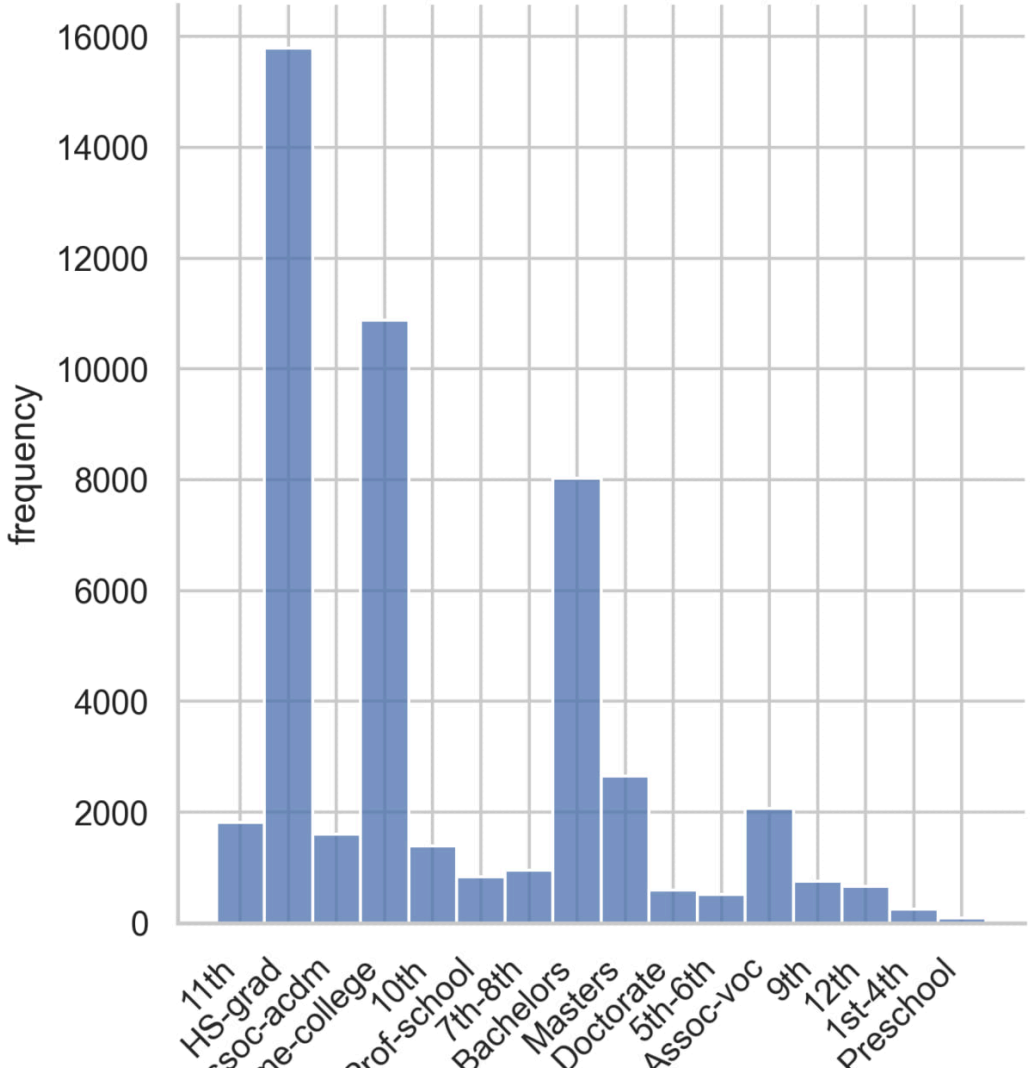



- Dominacja kategorii „Prywatne”: Kategoria „Prywatne” jest reprezentowana na prawie wszystkich poziomach edukacji, co wskazuje na jej dominację wśród innych kategorii zatrudnienia.

- Rzadkość innych kategorii zatrudnienia: Kategorie takie jak „Bez wynagrodzenia”, „Nigdy nie pracował”, „Federal-gov” i „Self-emp-inc” mają niewiele danych lub nie mają ich wcale na wykresie, co wskazuje na ich rzadkość w tym zbiorze danych.

- Zależność między wykształceniem a zatrudnieniem: Poziomy wykształcenia, takie jak „HS-grad” , „Some-college” i „Bachelors” obejmują większą różnorodność kategorii zatrudnienia w porównaniu z niższymi poziomami wykształcenia, takimi jak „Preschool” lub „1st-4th”.

  Częstotliwość wykształcenia

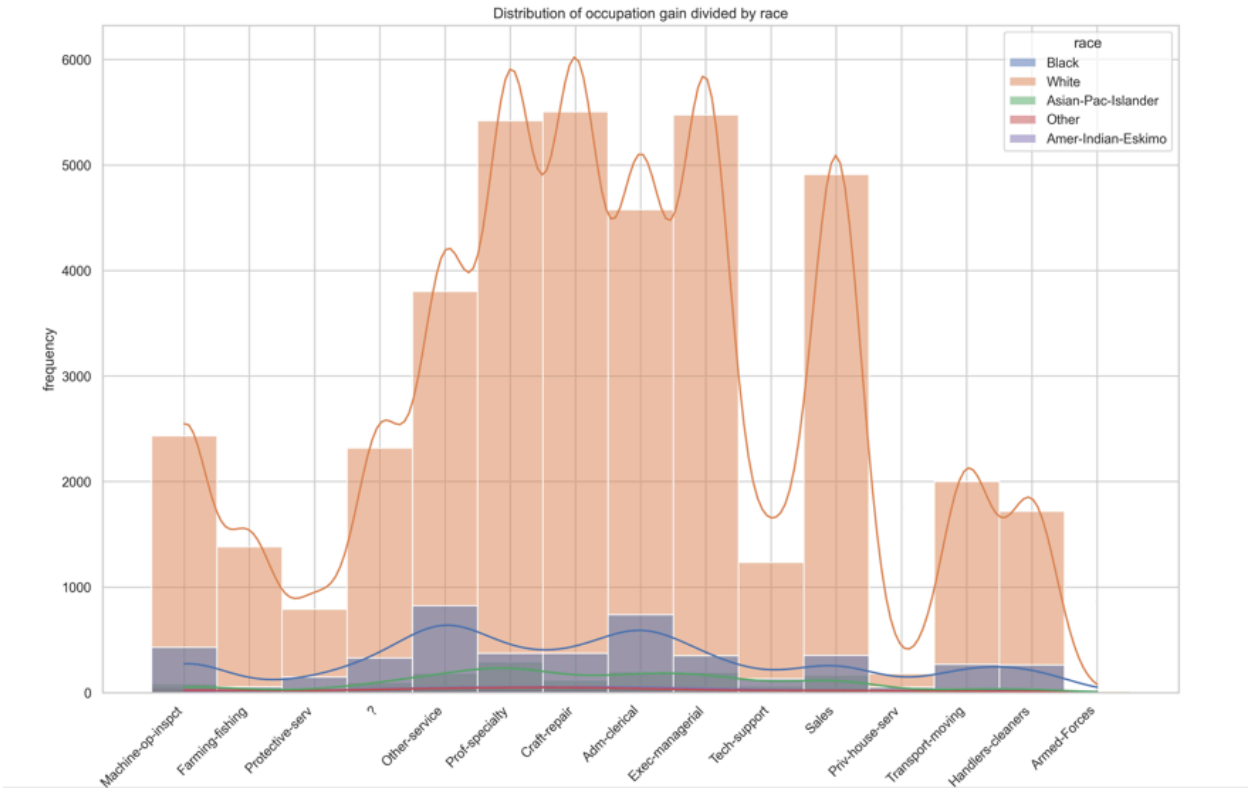



- Najczęstsze poziomy wykształcenia: Poziomy wykształcenia „HS-grad” i „Some-college” mają najwyższą częstotliwość, przekraczającą 10 000, z „HS-grad” na pierwszym miejscu.

- Niższa częstotliwość wykształcenia wyższego: Podczas gdy poziom „Bachelors” jest znaczący, częstotliwość dla kategorii „Masters” i „Doctorate” jest znacznie niższa, co może odzwierciedlać mniejszą liczbę osób z tymi poziomami wykształcenia.

- Bardzo niska częstotliwość wykształcenia podstawowego: Poziomy wykształcenia „Przedszkole” i „1-4” mają bardzo niską częstotliwość, co wskazuje, że są one niezwykle rzadkie w tym zbiorze danych.

Częstotliwość zawodów



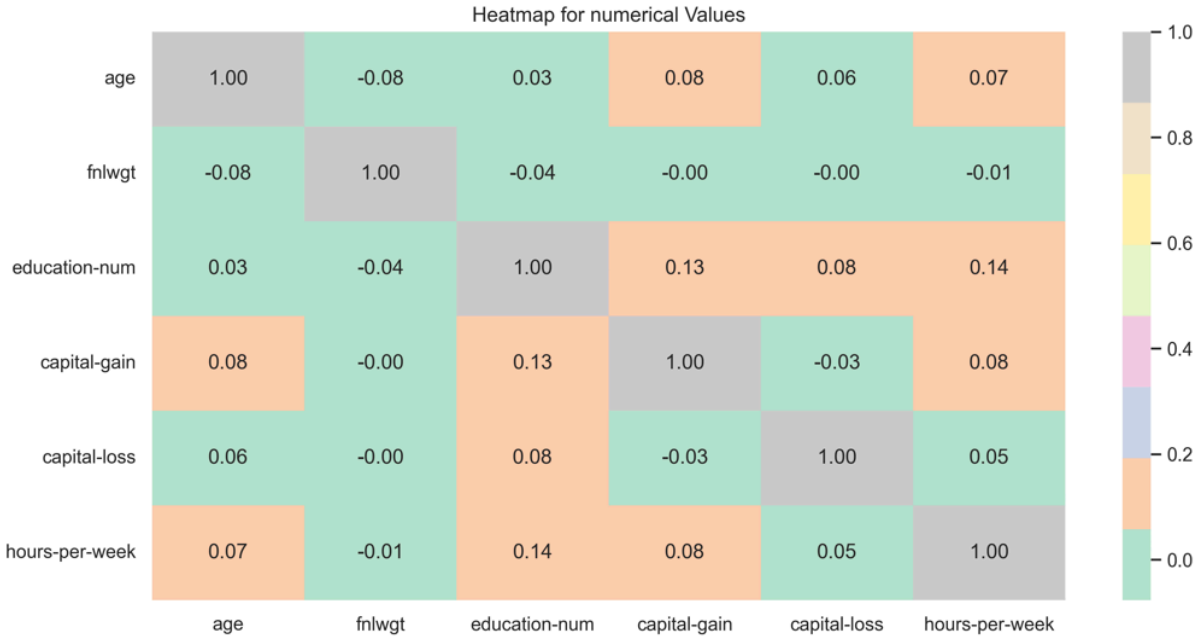

- Przewaga grupy białej

- Najpopularniejsze zawody dla białych to: wsparcie techniczne, rzemiosło-naprawy, specjalizacja zawodowa
-
- Najpopularniejsze zawody wśród Afroamerykanów to: administracja, usługi biurowe i inne usługi

- Najmniej popularne zawody to: rolnictwo-rybołówstwo, siły zbrojne, usługi domowe

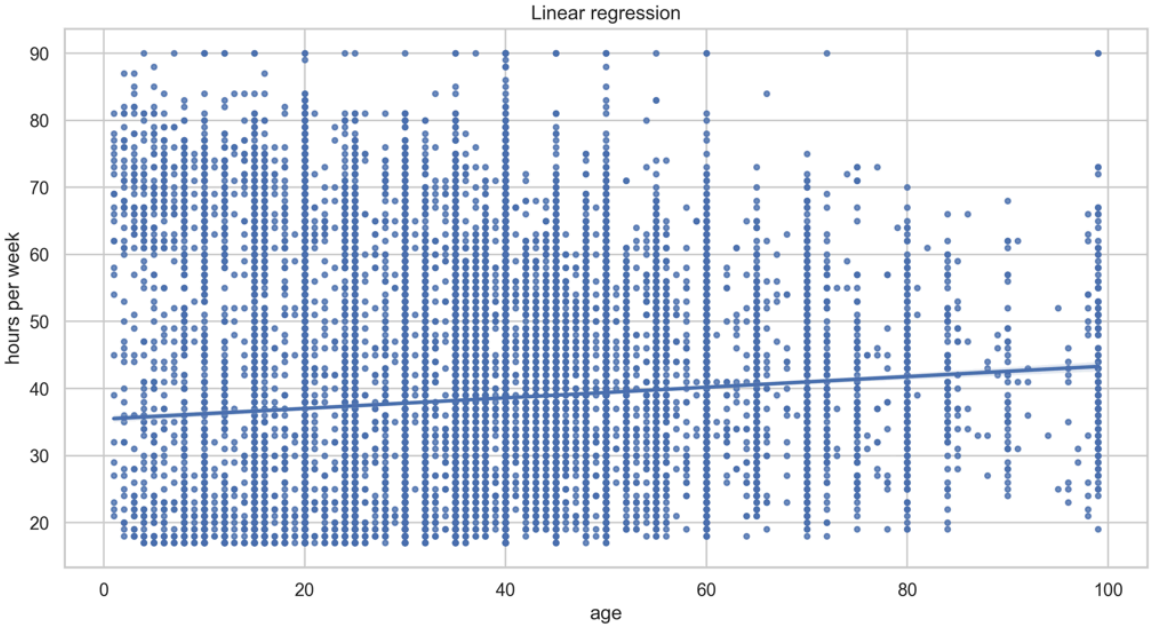
- Rzadkość poszczególnych zawodów: Niektóre kategorie, takie jak siły zbrojne, usługi prywatne i sprzątaczki, mają niską częstotliwość, niezależnie od grupy rasowej.

Mapa termiczna




Regresja liniowa

- Korelacje między zmiennymi są słabe (od -0,08 do 0,14), co wskazuje, że nie ma wyraźnych zależności liniowych. Oznacza to, że do przewidywania, na przykład dochodu, mogą być potrzebne dodatkowe czynniki lub modele nieliniowe.




- Wykres pokazuje słabą dodatnią zależność między wiekiem a liczbą godzin pracy w tygodniu. Regresja liniowa ma niewielkie nachylenie, co wskazuje na nieistotny wpływ wieku na liczbę godzin pracy. Dane są bardzo rozproszone, co wskazuje na dużą zmienność.

Pozostałe wykresy, które nie zostały uwzględnione w raporcie, znajdują się na githubie.