**Zadanie 1 (VEGA - język interaktywnych wizualizacji).** Używając pakietu “Vega-Altair”

* zilustruj *interakcje między prestiżem, zarobkami, edukacją różnych zawodów*, używając danych z badania “Duncan, O. D. (1961) A socioeconomic index for all occupations”
* odpowiedz na pytanie: która profesja cieszy się wysokim prestiżem pomimo niskich zarobków? 🙂

Wskazówka: można zaadaptować poniższy kod

import statsmodels.api as sm

from statsmodels.formula.api import ols

import altair as alt

from vega\_datasets import data

data = sm.datasets.get\_rdataset("Duncan", "carData", cache=True).data # using statsmodels, load data from the R collection

alt.Chart(data).mark\_circle().encode(

alt.X(alt.repeat("row"), type='quantitative'),

alt.Y(alt.repeat("column"), type='quantitative'),

color=# color by job type

tooltip= # show job name on hover

).properties(

width=150,

height=150

).repeat(

row=# income, education, prestige,

column=# income, education, prestige,

).interactive()

[Dokumentacja Vega-Altair](https://altair-viz.github.io/gallery/scatter_matrix.html)

[Oczekiwany wynik](https://vega.github.io/editor/#/url/vega--pJCFUADoEITGKR6qGHP63R7VEHMCH8lpCZggyxQxVOqG2KGAFaUCJ5IlSwim0BpImFYplSrVWoNJotNodHr9IYjE6GREA7LROaqNoUIiskE3Zz-GBOQI5QYAJisNjs4oMyF6g0irncYN6i7k-ewgKHuSUcF24heoLmLggs+487BnWXIFCa4cG862+n0T3KC9BIR4gP8J6DqaIBTO6GYypOWL3o+JhrEEL6lCuYTrig3RvlEMQeNhbAgWBp6QaopBMHs8F3jOc4oTkpToe+q7hBu3TCnhAHID+xEDkCczupguhwG41HnrRT70WCgwsBhn4SluO7-rEhG8eB-G5MJUAiaq+w3lOiF0WYYKLhUcmsYBsl-vh+6dGppE0Q+0QlnpGK3uJD5GahKAyeZWHIN0EhKTZ2HAZwx4ObkdCYASon6QhEnIcZgHCh+FnIBUHDBVx7H2RBfzKHsuliSA2DJJJyXIIMnR+V+gE1dZXGlIe4WgXxZ7vlAEZQJgaolYZFXedxb5pf5imNSpm69HlGkgKQUDRfIRLTH18WOUhz6sCNLH+Z0VmcZNpR2a1JH5bkWiqNgZyrW5BmJZt3G-qNdUBb02WTb03QzR1SiMGclBwGAcD4v191SSgP61RK3QNQdBEBcdEVnaVRwwFMmChtgewra5SrTp5g05CwZnMZhL29Ptu6HV9J3tZBqojjyuPufjG3g8gLBvaT8lsSTcMvmFSOzQQq6yiVZVeTkkPc+lEiU8p8MSNNtPqR1BCEPII5i2tHls5VpRBTL-ksEx-MQ4M32QWjMp7GcoYAjdeMDUlQ2DKlO0vaUv5m5uFSW3Mbh7BjkxZB4zN3QTLtE9tZMSiw8shRzpT+7k9T4jjoORw9nRc89Cm4VT8M4SnID1G80wrcojssxLhMmd0UPfgXCtghINNCx1kyXqGdjyPIWDh1itdR2CnONygjHvYrLAlyQlCkJrawY6oHhEOL5UjxPTF59+fOF6Pfsq5FCSyhjSxqrMOulRvD2Lg3Ruexx+8ChbR-I73KhENXU71Df7OmePZAi5DY+wkILNqqtILSg1oDYGBgr6-0lqPBqO9JD3x9gbEuAAlLBoYYrdV6oPOYiC66ARJqgoBXMfbVRLpMGC2k4q3SxCQzeyAJBPQ9hKCQzdE5HRLuRV4oYkSqAWCHEqLDb4xx5igRcPCuKLlfh3SCSgdC6BaEQ1Of9Krx0AYuU2z8gHtwgcfEMlBQxuCgKGdQqoFoTgQVooagpdHoIMZlEukBHihi0t-ZhDipaANhq45Ob9ZpuEUJgQgvcvEQFlGI+xSCtq6O9q4oxp1ZpEmhHAUMdB+4kg0aXPxYJ74UIqC4luAoiIhI6hrWw0xsm5J8cQwpMjyGcI3E4ia8MKjgLSZ3RmJAMYQBUJrbWTCmkJKAa02OG5FxP3KUAmeVTIJp3kOo8RzSMocOmTI5J8zZklxWTYxa6yJlDF0VQgxiNjHI0oIwCAGNbkDPyRI9mxS2kCkCfMnpdMBLlxOaQjKVkKGyKnmCIYJcSxpB7ssGY-zWEVCkbLOZidFHXNmtgdQ7o4ljM0aclB7ygH6PmRUQ+Si5iFnxISKi8SAX4u2UA5FXFlZkvPO8QGQkQY0tYXLQBLBPmJ2JiXDFGZnkbNzgS7hoKKnSGtNIIAA)

**Zadanie 2 (Związek prestiżu, edukacji i zarobków)** Używając danych z badania “Duncan, O. D. (1961) A socioeconomic index for all occupations”:

* zbuduj model objaśniający prestiż przez zarobki, edukację i rodzaj pracy (“niebieskie kołnierzyki”, “białe kołnierzyki itp).
* używając istotności współczynników (P>|t|), oceń dla których zawodach widać najmocniejszy wpływ edukacji na prestiż?

import statsmodels.api as sm

from statsmodels.formula.api import ols

# fit a model

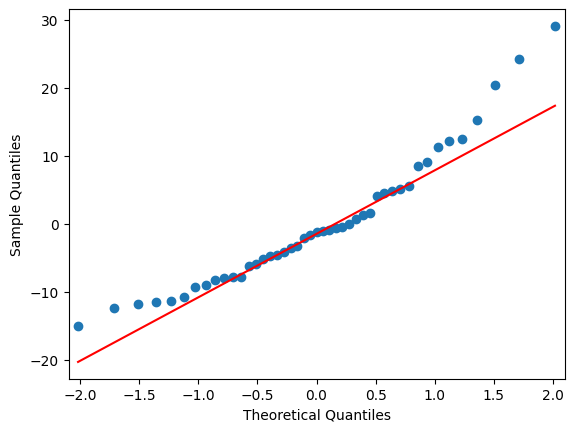
prestige = sm.datasets.get\_rdataset("Duncan", "carData", cache=True).data

model1 = ols("prestige ~ income + education:C(type)", data=prestige).fit() # NOTE: make the education coefficient depend on job type

model1.summary()

UWAGA: operatory “:” oraz “\*” służą do wprowadzania [interakcji](https://bkenkel.com/pdaps/specification.html#interaction-terms-1)

**Zadanie 3 (Diagnostyka i naprawa regresji).** Dla reszt modelu z poprzedniego zadania:

* czy reszty regresji mają rozkład normalny (wskazówka: <summary> raportuje test Jarque-Bera)
* narysuj wykres typu QQ dla reszt regresji; oczekiwany wynik to  
  

zidentyfikuj anomalię w prawym górnym rogu, któremu zawodowi odpowiada? (wskazówka: popatrz na wysokie wartości model1.resid ) 🙂

* dopasuj model ponownie usuwając anomalię z danych. jak wygląda teraz wynik testu Jarque-Bera?

import statsmodels.api as sm

from statsmodels.formula.api import ols

# fit a model

prestige = sm.datasets.get\_rdataset("Duncan", "carData", cache=True).data

mask = ~prestige.index.isin(["anomaly"])

model2 = ols("prestige ~ income + education:C(type)", data=prestige[mask]).fit() # NOTE: make the education coefficient depend on job type

model2.summary()

**Zadanie 4 (testowanie współczynników regresji)** Niezerowość / nietrywialność współczynników regresji testowana jest regułą wykorzystującą rozkład T:  
<https://bkenkel.com/pdaps/inference.html#single-variable-hypotheses>

Odnosząc się do tabeli poniżej (lub wyników innej regresji) pokaż jak wyliczyć T-value oraz P-value używając wartości “Coefficient” oraz “Std Error”.

| **Variable** | **Coefficient** | **Std Error** | **T-value** | **P-value** | **95% CI Lower** | **95% CI Upper** |
| --- | --- | --- | --- | --- | --- | --- |
| Intercept | -0.1078 | 6.013 | -0.018 | 0.986 | -12.260 | 12.044 |
| Income | 0.5908 | 0.093 | 6.336 | 0.000 | 0.402 | 0.779 |

**Zadanie 5 (Rozwiązanie regresji a pseudoodwrotność macierzy)** Znajdź związek między rozwiązaniem OLS

<https://bkenkel.com/pdaps/ols-matrix.html#the-ols-estimator>

a pseudo-odwrotnością macierzy X