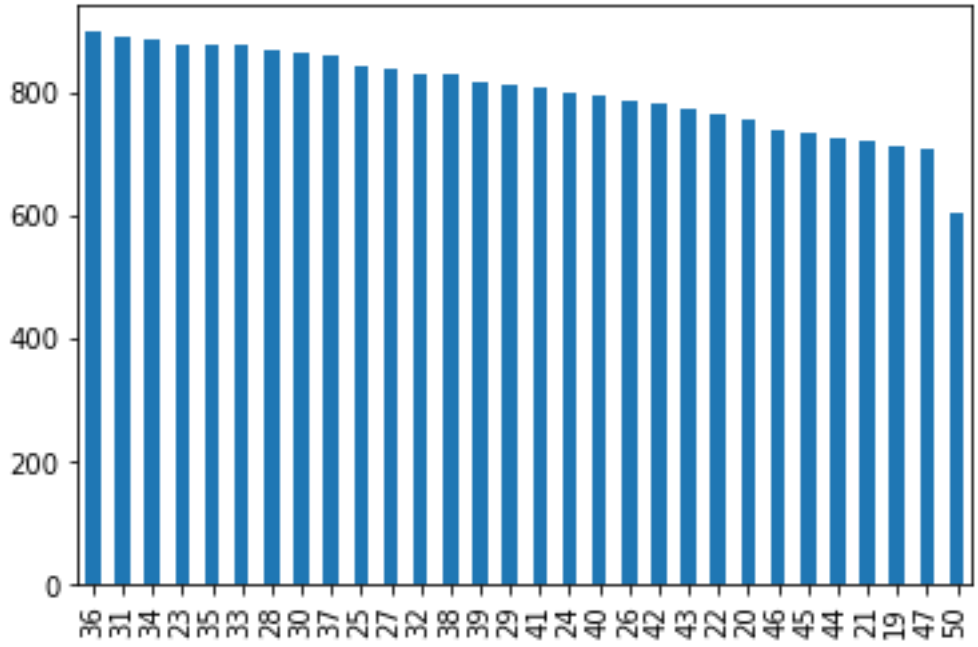
Desafio Cientista de Dados

Gustavo Cabral de Barros
(11) 99999 – 2125
gust.cbarros@gmail.com

1. Gráficos e dados

Abaixo estão os gráficos realizados dos dados do wage, e algumas estatísticas descritivas.

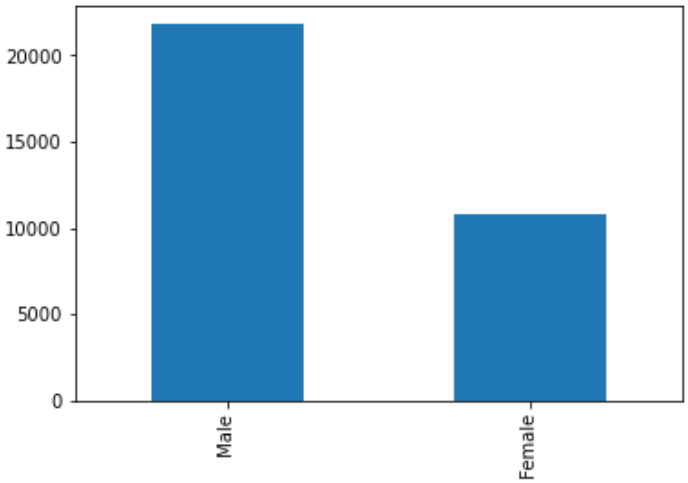
Distribuição das 30 idades com maior frequência



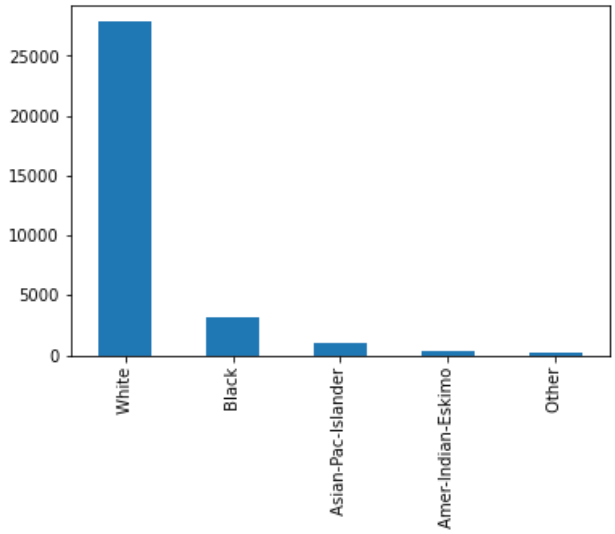
Distribuição da nação das pessoas pesquisadas:

United-States	29169
Mexico	643
?	583
Philippines	198
Germany	137
Canada	121
Puerto-Rico	114
El-Salvador	106
India	100
Cuba	95
England	90
Jamaica	81
South	80
China	75
Italy	73
Dominican-Republic	70
Vietnam	67
Guatemala	64
Japan	62
Poland	60
Columbia	59
Taiwan	51
Haiti	44
Iran	43
Portugal	37
Nicaragua	34
Peru	31
France	29
Greece	29
Ecuador	28
Ireland	24
Hong	20
Cambodia	19
Trinadad&Tobago	19
Laos	18
Thailand	18
Yugoslavia	16
Outlying-US(Guam-USVI-etc)	14
Honduras	13
Hungary	13
Scotland	12
Holand-Netherlands	1

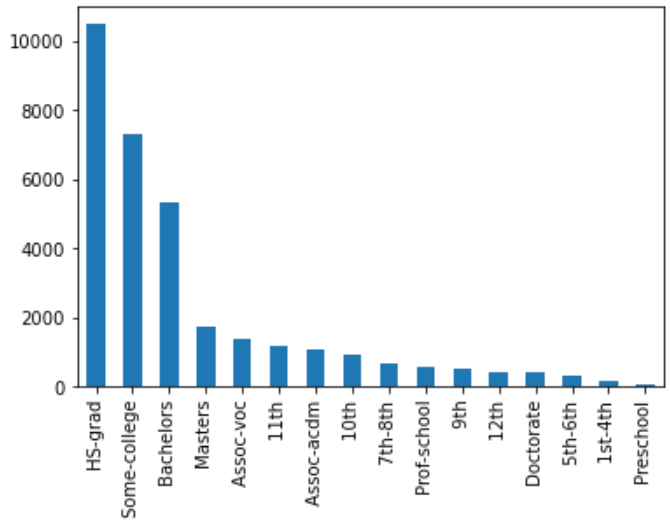
Distribuição dos homens e mulheres



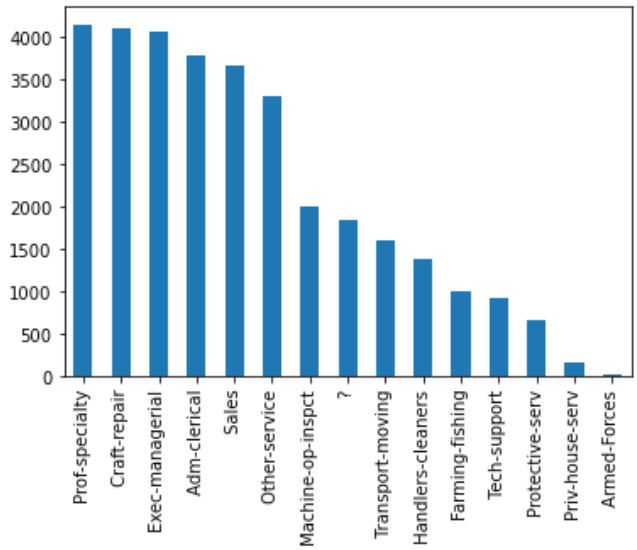
Distribuição do grau das “raças” ou etnias dos usuários:



Distribuição do grau de educação dos usuários:



Distribuição de qual a ocupação dos usuários:



Estatísticas descritivas dos dados quantitativos dos usuários

	Index	age	fnlwgt	education_num	capital_gain	capital_loss	hours_per_week
count	32560.000000	32560.000000	3.256000e+04	32560.000000	32560.000000	32560.000000	32560.000000
mean	16279.500000	38.581634	1.897818e+05	10.080590	1077.615172	87.306511	40.437469
std	9399.406719	13.640642	1.055498e+05	2.572709	7385.402999	402.966116	12.347618
min	0.000000	17.000000	1.228500e+04	1.000000	0.000000	0.000000	1.000000
25%	8139.750000	28.000000	1.178315e+05	9.000000	0.000000	0.000000	40.000000
50%	16279.500000	37.000000	1.783630e+05	10.000000	0.000000	0.000000	40.000000
75%	24419.250000	48.000000	2.370545e+05	12.000000	0.000000	0.000000	45.000000
max	32559.000000	90.000000	1.484705e+06	16.000000	99999.000000	4356.000000	99.000000

Os dados escolhidos, são o que achei mais interessante de trazer. Ver como é desigual a distribuição dos países, sendo majoritariamente pessoas dos Estados Unidos pesquisados.

O Homem Branco é o mais comum dos sexos e etnias, mas é importante também destacar as demais etnias e as mulheres das pessoas pesquisadas, por isso trazer esses dois gráficos.

Além disso saber qual o grau de educação das pessoas é importante para ter o conhecimento de que entre as pessoas pesquisadas, majoritariamente elas possuem o ensino de base.

A idade é bem importante na definição do salário anual, pois existem “ranges” de idade nas quais as pessoas ganham mais ou menos. As profissões estão relacionadas com a idade, e definem os possíveis salários por setor.

Por fim, coloquei as estatísticas descritivas das variáveis quantitativas para poder ter uma noção de qual é a média, valores máximos e mínimos entre outros dados importantes.

2 – Para a realização da previsão dos salários, eu realizei uma regressão, utilizando os mínimos quadrados ordinários (MQO), pois os dados não dependem de um fator temporal. Abaixo está o resultado da regressão:

OLS Regression Results

Dep. Variable:	yearly_wage	R-squared:	0.342
Model:	OLS	Adj. R-squared:	0.342
Method:	Least Squares	F-statistic:	1236.
Date:	Sat, 23 Jul 2022	Prob (F-statistic):	0.00
Time:	17:17:34	Log-Likelihood:	-11401.
No. Observations:	30724	AIC:	2.283e+04
Df Residuals:	30709	BIC:	2.296e+04
Df Model:	14		
Covariance Type:	HC1		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.5836	0.012	-50.561	0.000	-0.606	-0.561
age	0.0037	0.000	21.337	0.000	0.003	0.004
workclass	-0.0122	0.003	-3.584	0.000	-0.019	-0.006
fnlwgt	8.782e-08	1.88e-08	4.671	0.000	5.1e-08	1.25e-07
education	-0.0104	0.004	-2.791	0.005	-0.018	-0.003
education_num	0.0416	0.001	39.661	0.000	0.040	0.044
marital_status	0.2516	0.006	45.432	0.000	0.241	0.262
occupation	0.0755	0.004	18.605	0.000	0.068	0.083
relationship	-0.0455	0.004	-11.229	0.000	-0.053	-0.038
race	-0.0267	0.005	-5.182	0.000	-0.037	-0.017
sex	-0.0272	0.004	-6.263	0.000	-0.036	-0.019
capital_gain	8.349e-06	3.25e-07	25.702	0.000	7.71e-06	8.99e-06
capital_loss	0.0001	5.59e-06	17.988	0.000	8.97e-05	0.000
hours_per_week	0.0030	0.000	16.388	0.000	0.003	0.003
native_country	-0.0160	0.006	-2.870	0.004	-0.027	-0.005

Omnibus:	1607.769	Durbin-Watson:	2.001
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1695.242
Skew:	0.544	Prob(JB):	0.00
Kurtosis:	2.627	Cond. No.	1.39e+06

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)
[2] The condition number is large, 1.39e+06. This might indicate that there are strong multicollinearity or other numerical problems.

Todas as variáveis são estatisticamente diferentes de zero, como podemos ver pelos P-valores. Pela quantidade de dados, a leis dos grandes números, garante a normalidade dos resíduos. Apliquei o método de correção dos erros robustos de White, para poder corrigir o problema da heteroscedasticidade.

Ao pesquisar esse problema, vi que a solução era feita a partir de Machine Learning, mas por não ter tanto conhecimento nessa parte, quis fazer pelos métodos econométricos de regressão, que também é o que eu gostaria de me especializar. Sei que o R^2 não foi muito alto, apesar de ser um bom valor para dados em cross-section, mas com o tempo que eu tive e os meus conhecimentos na área, resolvi estimar pelo MQO.

O ideal seria realizar diversos testes, modificando os parâmetros adicionados e utilizar critérios como o de Jarque-Bera, Critério de informação de Akaike (AIC) e Critério Bayesiano de Schwarz (BIC), para decidir o melhor modelo, porém pela limitação de tempo que tive, acabei optando por utilizar todas as variáveis. Tenho conhecimento que para a formação de salários, as variáveis que mais influenciam nos seus valores, pelas minhas pesquisas, acredito que as mais importantes são a idade, profissão, país que vive, raça/etnia, sexo. Dentre as outras variáveis, acredito que algumas ou todas são importantes também, mas não possuo conhecimento o

suficiente nessa área para saber quais retirar ou manter, por conta disso mantive todas para não cometer o erro de retirar algo importante.

Por fim, para poder realizar a regressão, transformei os valores qualitativos em numéricos, e agrupei os dados de acordo com a proximidade que eles possuem. Na planilha, que está junto desse arquivo, está a forma que dividi, mas estou colocando abaixo também para deixar informado.

age	workclass	workclass	fnlwgt	education	education	education	marital_status	marital_st
Quantitativo	Self-emp-not-inc	2	Quantitativo	Bachelors	1	Quantitati	Married-civ-spouse	1
	Private	0		HS-grad	1		Divorced	0
	State-gov	1		11th	0		Married-spouse-absent	1
	Federal-gov	1		Masters	2		Never-married	0
	Local-gov	1		9th	0		Separated	0
	?	1		Some-college	2		Married-AF-spouse	1
	Self-emp-inc	2		Assoc-acdm	1		Widowed	0
	Without-pay	2		Assoc-voc	1			
	Never-worked	2		7th-8th	0			
				Doctorate	2			
				Prof-school	1			
				5th-6th	0			
				10th	0			
				1st-4th	0			
				Preschool	0			
				12th	0			

occupation	occupatio	relationship	relationsh	race	race (num	sex	capital_gain	capital_loss	hours_per_week
Exec-managerial	1	Husband	0	White	0	Male	Quantitativo	Quantitativo	Quantitativo
Handlers-cleaners	0	Not-in-family	1	Black	1	Female			
Prof-specialty	1	Wife	0	Asian-Pac	1				
Other-service	0	Own-child	0	Amer-Indi	1				
Adm-clerical	0	Unmarried	1	Other	1				
Sales	1	Other-relative	0						
Craft-repair	0								
Transport-moving	0								
Farming-fishing	0								
Machine-op-inspct	1								
Tech-support	1								
?	1								
Protective-serv	2								
Armed-Forces	2								
Priv-house-serv	0								

native_co	native_country (number)	
Canada	0	High income
England	0	High income
France	0	High income
Germany	0	High income
Greece	0	High income
Holand-N	0	High income
Hong	0	High income
Hungary	0	High income
Ireland	0	High income
Italy	0	High income
Japan	0	High income
Outlying-U	0	High income
Poland	0	High income
Portugal	0	High income
Puerto-Ri	0	High income
Scotland	0	High income
Taiwan	0	High income
Trinadad&	0	High income
United-St	0	High income
Yugoslavia	0	High income
Cambodia	2	Lower middle income
El-Salvado	2	Lower middle income
Haiti	2	Lower middle income
Honduras	2	Lower middle income
India	2	Lower middle income
Iran	2	Lower middle income
Laos	2	Lower middle income
Nicaragua	2	Lower middle income
Philippine	2	Lower middle income
Vietnam	2	Lower middle income
China	1	Upper middle income
Columbia	1	Upper middle income
Cuba	1	Upper middle income
Dominican	1	Upper middle income
Ecuador	1	Upper middle income
Guatemal	1	Upper middle income
Jamaica	1	Upper middle income
Mexico	1	Upper middle income
Peru	1	Upper middle income
South	1	Upper middle income
Thailand	1	Upper middle income

Vale destacar que utilizei o critério de renda do Banco Mundial para dividir os países entre High income, Lower middle income e Upper middle income. E para as etnias, como o branco tem acesso mais fácil a empregos com salários maiores, não

enfrentam preconceitos, decidi colocar todas as minorias como o mesmo valor, pois elas enfrentam dificuldades parecidas que influenciam nos valores dos seus salários.

Gostei muito de realizar esse desafio, e por conta de eu estar no final do semestre na universidade com muitas provas e trabalhos, acabei não conseguindo realizar tudo o que eu gostaria, para entregar um desafio ainda melhor. Como por exemplo, ter feito todas as variáveis qualitativas serem vetores de Dummy com valor 0 ou 1, e rodar a regressão com essas dummies, para cada dados.