Analýza pracovných ponúk na pozíciu Data Analyst

Adrián Pauer

Letný semester 2022/2023

Obsah

1 Úvod do témy

Tému pracovných ponúk som si vybral z dôvodu, že ma zaujíma pracovný trh a oblasti, v ktorých by som potenciálne vedel získať uplatnenie. Cieľom projektu je preskúmať základné trendy v dátach, skúmať oblasti a sektory, v ktorých sú ponuky prezentované. Význam projektu spočíva v spracovaní dát a tvorbe grafov, ktoré umožňujú všímať si rôzne súvislosti. V projekte sa používajú rôzne programovacie jazyky ako R, Python, SQL. Práca sa zameriava na pozorvanie javov ako sú počet ponúk v sektoroch, skúmanie platov, taktiež aj korelácia medzi vybranými dátami či spracovanie textu. Implementácia je rozdelená do viacerých súborov.

2 Popis Dát

Dáta boli stiahnuté zo stránky Kaggle [1]. Následne boli uložené na server, kde som s nimi pracoval. Dataset má celkovo 2253 riadkov a obsahuje 15 stĺpcov. Jednotilvé stĺpce popisujú ponuky na pozíciu Data Analyst v USA. Podľa chýbajúcich hodnôt som sa rozhodol, ktoré stĺpce z analýzy vyradím. Chýbajúce hodnoty v jednolivých stĺpcoch vyjadruje tabuľka.

- `title` 0 0.0%
- `salary` 1 0.04%
- `description` 0 0.0%
- `rating` 272 12.07%
- `company` 0 0.0%
- `location` 0 0.0%
- `headquarters` 172 7.63%
- `size` 163 7.23%
- `founded` 660 29.29%
- `ownership` 163 7.23%
- `industry` 353 15.67%
- `sector` 353 15.67%
- `competitors` 1732 76.88%
- `revenue` 163 7.23%
- `easyApply` 2173 96.45%

Stĺpce s vysokým percentom chýbajúcich hodnôt neboli nepoužité. Sú to `easyApply` a `competitors`.

3 Analýza pomocou SQL

Táto časť spočíva vo využití príkazov ako `GROUP BY` a `ORDER BY`, ktoré sú použité na tvorbu základných pozorovaní. Implentácia sa nachádza v súbore `observationSQL.py`. Výsledné tabuľky boli uložené do CSV súboru, z ktorých som následne v Rstudio vykreslil príslušné grafy. Vygenerované súbory sú uložené v priečinku `generatedCSVs` a na tvorbu `.csv` slúži skript `toCSV.py`.

3.1 20 Najžiadanejších sektorov

Pre jednotlivé sektory som spočítal, koľko ponúk celkovo prislúcha jednotlivým sektorom.

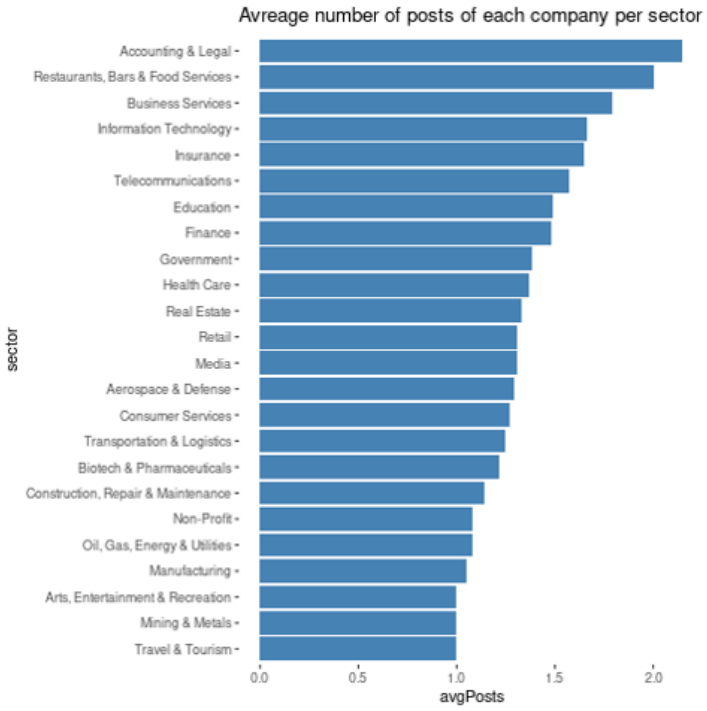
Information Technology	570
Business Services	524
Finance	169
Health Care	151
Education	52
Insurance	51
Accounting & Legal	43
Media	42
Manufacturing	40
Retail	38
Government	36
Biotech & Pharmaceuticals	33
Non-Profit	26
Aerospace & Defense	22

Vidíme, že najviac ponúk bolo publikovaných v sektore `Inforamtion Technology`, čo je celkom očakávaný výsledok. Prekvapivo sú v piatich najžiadanejších aj sektory ako `Health Care`, či `Education`, v ktorých by som nečakal až tak veľa pracovných ponúk tohto typu. Možnou príčinou je aj dnešný rovoj bioinformatiky a oblastí s ňou spojených, kde sa využívajú poznatky dátovej analýzy vo veľkej miere.

3.2 Priemerný počet ponúk firiem v sektoroch

Zaujímalo ma, aký je počet ponúk na firmu v jednotlivých sektorov. Intuitívne a aj na základe výsledkov predošlej časti očakávam, že v informatických sektorov bude najviac ponúk.

sektor	priemerný počet ponúk
accounting & Legal	2.15
Restaurants, Bars & Food Services	2.0
Business Services	1.79
Information Technology	1.66
Insurance	1.65
Telecommunications	1.57
Education	1.49
Finance	1.48
Government	1.38
Health Care	1.37
Real Estate	1.33
Retail	1.31
Media	1.31
Aerospace & Defense	1.29
Consumer Services	1.27
Transportation & Logistics	1.25
Biotech & Pharmaceuticals	1.22
Construction, Repair & Maintenance	1.14

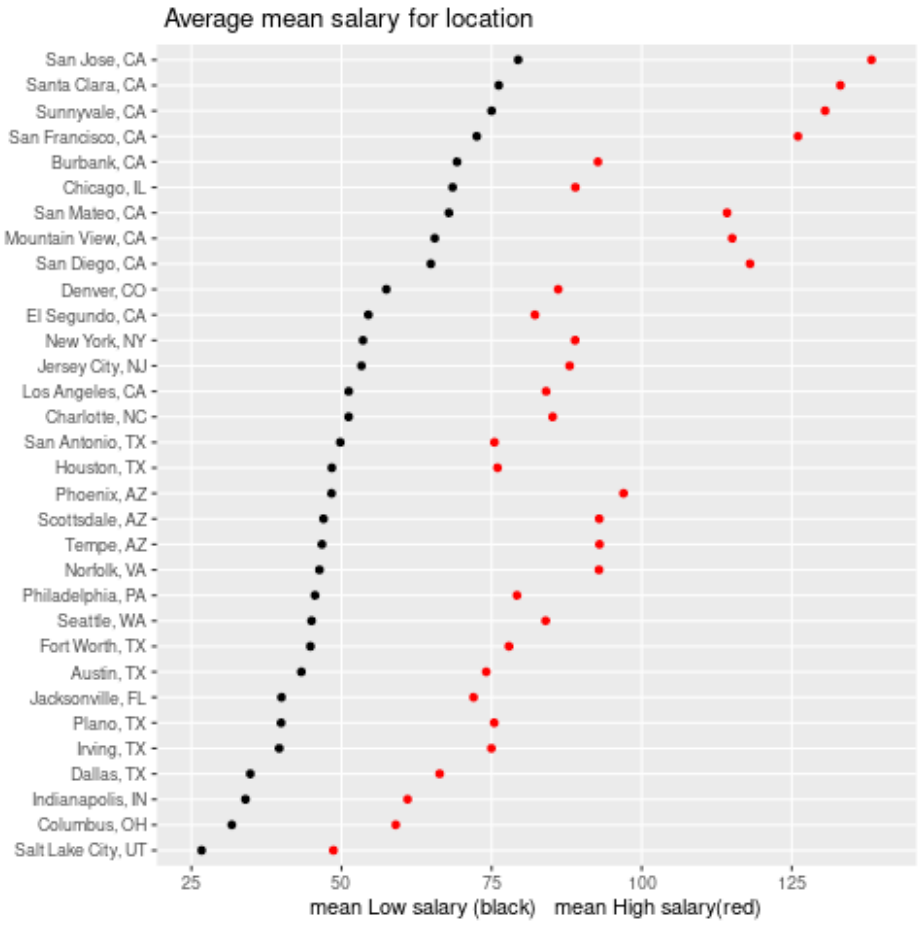


Obr. 1: Prvých 14 sektorov, do ktorých posielajú firmy priemerne najviac ponúk

Najväčší počet ponúk dosiahli sektory `accounting and Legal` a `Restaurants, Bars & Food Services`. Výsledok sa teda nezhoduje s mojimi predpokladmi. Po hlbšom preskúmaní dát som zistil, že firma Andiamo zverejnila celkovo tri ponuky, všetky pre sektor `Restaurants, Bars and Food Services`. Teda takéto prípady skresľujú výsledok.

3.3 Plat v jednotlivých oblastiach

Stĺpec `salary` som rozdelil na dva samostatné, `low` a `high`. Následne som vybral oblasti ktoré mali aspoň 15 ponúk a pozrel sa na priemerné hodnoty minimálneho a maximálneho ponúkaného platu.



Obr. 2: Platy v oblastiach

Graf znázorňuje na horizontálnej osi veľkosť priemerného minimálenho platu čiernou farbou a veľkosť priemerného maximálneho platu červenou farbou. Hodnoty sú zoradené podľa minimálneho platu a sú vyjadrené v tisíckach dolárov na rok . Najvyšší plat ponúka mesto San Jose, ktoré sa nachádza v Kaliforni. Kedže v tejto oblasti je priemysel rozvinutý, v popredných priečkach sú hlavne mestá z tohto štátu. Ponuku platu môže tiež zvyšovať aj fakt, že tento štát je prímorský, teda aj turizmus prispieva k bohatstvu oblasti.

Prekvapujúce je, že Chicago ponúka maximálny ročný plat v priemere okolo 85000 dolárov, čo je porovnateľné s mestami, ktoré ponúkajú menší minimálny plat.

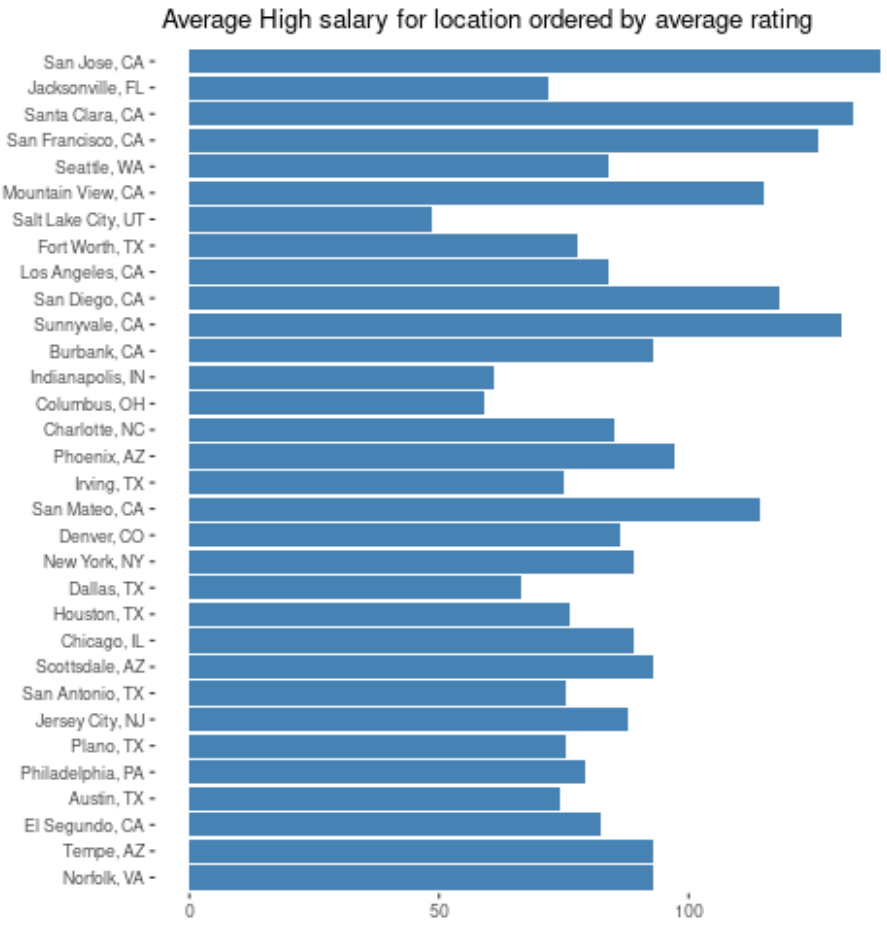
Vo výsledkoch najmenší priemerný plat ponúkajú mestá ako `Columbus` a `Salt Lake City`. Väčšina z nich leží vo vnútrozemí, aj to môže byť príčinou.

4 Analýza pomocou R

V tejto časti používam upravené dáta, ktoré som po očistení uložil do súboru `jobs.csv` v priečinku `generatedCSVs`. Práca spočíva vo využití knižnice `ggplot2` a tvorbe príslušných grafov. Celkové spracovanie a implementácia sa nachádza v súbore `graphs.R`. Skúmal som faktory ako `hodnotenie` a `veľkosť firmy`.

4.1 Vplyv hodnotenia firiem na priemerný plat

Pozrime sa teraz, akú zmenu v poradí miest zapríčiní zoradenie dát podľa priemerného hodnotenia v danom meste. Predpokladáme, že výsledky sa nebudú výrazne líšiť, aj keď nevylučujeme možnosť, kedy aj mesto s menším ponúkaným platom má vyššie hodnotenie. Pozorovania sú znázornené na nasledujúcom grafe. Uvažovali sme priemený maximálny plat.



Obr. 3: Platy v oblastiach

Mesto `San Jose` splnilo naše predpoklady. Ak sa ale pozrieme na mestá ako `Chicago` alebo `New York`, sú umiestnené v strede. Obidve mestá majú viac ako sto zverejnených ponúk, kdežto `Salt Lake City` má iba dvatsaťdva a je umiestné ako siedme. Preto mestá s nenej ponukami majú v niektorých prípadoch lepšie hodnotenie ako mestá s veľa ponukami.
Z pohľadu výšky priemerného maximálneho platu a hodnotenia by som určil tri najlepšie mestá `San Jose`, `Santa Clara`, `San Francisco`

4.2 Čím staršia firma, tým vačší plat

Zaoberal som sa otázkou, či má rok založenia firmy vplyv na ponuku maximálneho platu. V tejto analýze som použil dáta, kde sa v skúmaných stĺpcoch nenachádzajú chýbajúce hodnoty. Skúmaný dataset mal 1500 riadkov.

Obr. 4: Výška platu vzhľadom na rok založenia

Graf znázorňuje na horizontálnej osi rok založenia firmy. Hodnoty vetikálnej osi znázorňujú ponúkaný plat.

Na grafe môžme pozorovať, že najstaršia firma ponúka priemerný maximálny plat. Ostatné staršie firmy ponúkajú menej alebo tiež priemerne veľa. Najvyššie ponúkané platy sa vyskytujú pre firmy založené okolo roku 2000 ale aj 1900.

Z obrázku môžme taktiež vidieť, že staršie firmy majú vačšinou priemerné až podpriemerné hodnotenie.

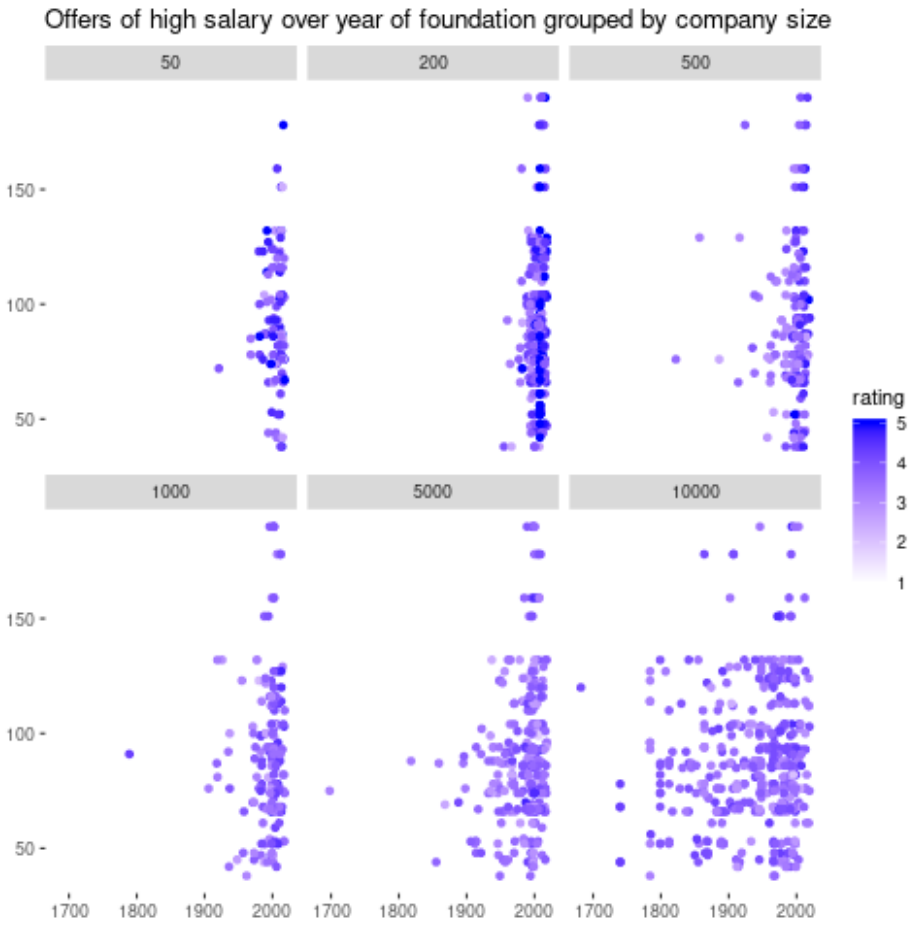
Teda usudzujem, že závislosť medzi rokom založenia firmy a ponuky maximálneho platu nebude. Potvrdzuje to aj nasledovné skúmanie korelácie medzi týmito dvomi premennými.

Pearson's product-moment correlation
t = 3.8772, df = 1570, p-value = 0.00011
alternative hypothesis: true correlation is not equal to 0
sample estimates: cor 0.09738601

Výsledná P-hodnota vyšla veľmi malá, čo znamená že alternatívnu hypotézu zamietame. Teda korelácia medzi týmito premennými existuje, čo je dosť prekvapujúce. Korelačný koeficient je ale veľmi malý.

4.3 Rozdelenie ponúk podľa počtu zamestnancov

Pozrel som sa, ako sa správajú ponuky platov vzhľadom na počet zamestnancov.



Obr. 5: Výška platu vzhľadom na rok založenia podľa počtu zamestnancov

Najvyššie ponúkané platy sa nachádzajú v každej skupine. Platy vo veľkých firmách sú skôr priemerné až podpriemerné.

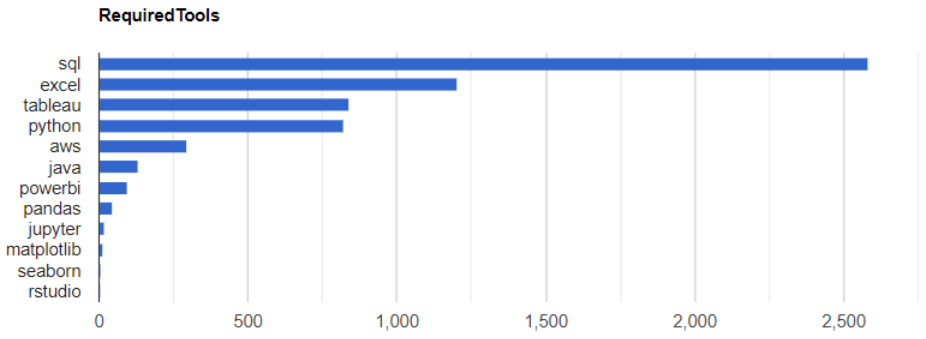
Zaujímavé je, že hodnotenie firiem s menším počtom zamestancov je najvyššie. Taktiež ich rok založenia je okolo 2000. Teda na trhu sa darí novým a menším firmám. Aj dnešná doba potvrdzuje, že sa na trhu dominujú tzv. startup-y a v mnohých prípadoch ponúkajú vyšší plat.

5 Spracovanie textu a clustering

Analyzoval som stĺpec `description`, kde som sa snažil odhaliť požadované zručnosti, vzdelanie a vlastnosti na uchádzačov o jednotlivé ponuky. Pracoval som s knižnicou `Sklearn` a nástrojmi ako `Vectorizer` a `k-means`. Na vykreslenie jednotlivých skupín som použil knižnicu `WordCloud` a `Matplotlib`. Implenetácia sa nachádza v priečinku `Vectorizer`, skript `getRequirements.py`.

5.1 Požadované nástroje, vlastnosti a vzdelanie

Text zo stĺpca `description` som normalizoval, odránil stop slová a spočítal výskyty jednotlivých slov. Výsledky som uložil do SQL tabuliek a následne zobrazil pomocou `javascript` ako `Bar-chart`.

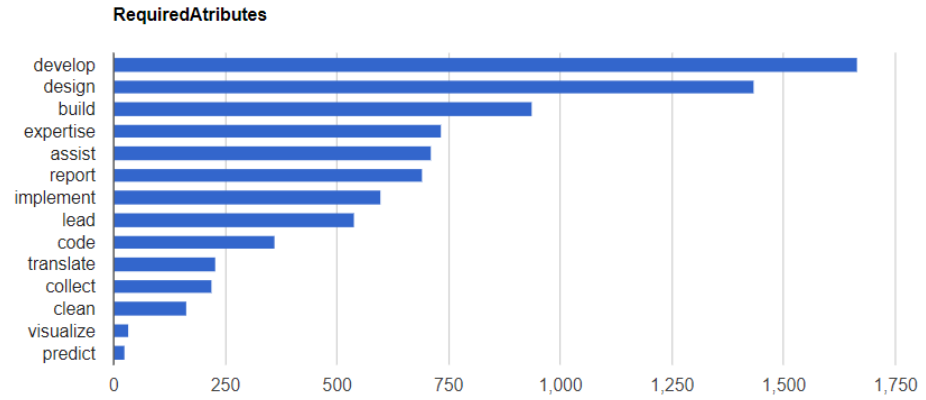


Obr. 6: Požadované nástroje

Najžiadanejší nástroj na prácu je `SQL`, ktoré je v ponukách zmienené okolo 2500 krát. Vidíme, že zastúpenie slova SQL v texte je dosť výrazné vzhľadom na ostatné skúmané nástroje.

Prekvapivé je, že `Excel` je druhý najžiadanejší. Dobrý námet na skúmanie je teda otázka, či sa patrí `Excel` v USA k najpoužívanejším nástrojom.

`Rstudio` je zaznamenané iba tri krát, príčinou je, že po vektorizácii textu sa odstránia aj samotatné písmená textu. Na pozíciu Data Analyst sú žiadané aj populárne programovacie nástroje ako `Python`, `Amazon web services` či `Java`.



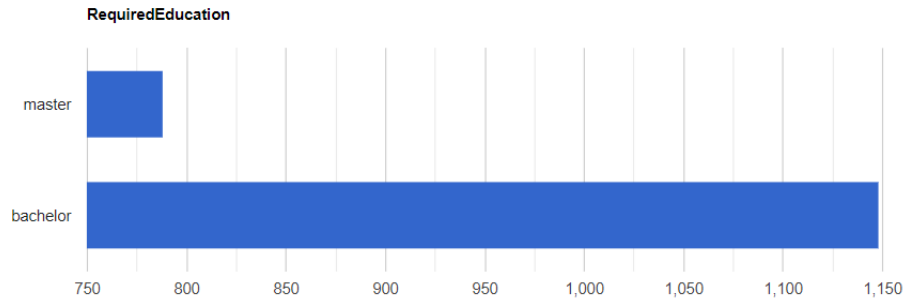
Obr. 7: Požadované vlastnosti

Výrazne veľa výskytov zaznamenali vlastnosti `develop` a `design`, ktoré sú typické pre pozície v oblasti dátovej

vedy.

Medzi najčastejšími sú aj `extertise` a `build`. Z toho usudzujem, že pracovná náplň v mnohých prípadoch zahŕňa aj výskum a samostatnú tvorbu.

Vlastnosť `predict` má najmenej výskytov, čo je dosť prekvapujúce, keďže pri umelej inteligencii často používame prediktívne modely.



Obr. 8: Požadované vzdelanie

V ponukách silno dominuje požiadavka na titul bakalára.

5.2 Clustering

V tejto časti som použil `Tfidf Vectorizer`, ktorý slovám priradzuje dôležitosť v danom texte. Použitím metódy `k-means` som vytvoril skupiny slov.

Obr. 9: Skupiny

Každá skupina obsahuje niektoré slová rovnaké. `Cluter 0` zhromažďuje slová ako práca, zručnosti skúsenosti, tím. Povedal by som, že zachytáva prevažne požiadavky na uchádzača.

V `Cluter 1` sú slová ako reports, develop, sql, solutions. Teda popisujú, čo budú uchádzači robiť.

Myslím, že `Cluter 2` je veľmi podobný ako `Cluter 1`. Rozdiel vidím v tom, že slová v `Cluter 2` popisujú viac hardvérové a technickejšie objekty ako packages, computer, database, systems kdežto v `Cluter 1` sú prevažne abstraktnejšie vlastnosti.

6 Zobrazenie výsledkov

Výsledky analýzy ako aj príslušné grafy sú zobrazené cez `Flask` s použitím `Javascript` pre vykreslenie niektorých grafov. Implementácia sa nachádza v priečinku `simple_flask`.

7 Záver

Výsledky anaylýzy ukazujú, že najžiadanejšou oblasťou pre ponuku dátového analytika je informatická, čo vieme potvrdiť aj v súčasnosti.

Platovo najlepšia oblasť sa javí ako Kalifornia a jej mestá, čo potvrdilo aj pozorovanie, ktoré uvažovalo aj hodotenia firiem.

Skúmanie vplyvu roku založenia na ponuku platu vyvrátilo závislosť medzi týmito premennými. Výsledky ale ukázali, že firmy s malým počtom zamestnancov majú najlepšie hodnotenie.

Najžiadanejšie nástroje na prácu sú sql, tableau, excel a iné. Zamestnávatelia žiadajú od uchádzačov vlastnosti ako develop, design a prevažuje požiadavka na bakalárske vzdelanie.

V časti 3.2 sa nepodarilo presne odhadnúť sektory, do ktorých firmy poslali svoje ponuky. Analýzu skresľujú totiž prípady kedy firma poslala všetky svoje ponuky iba pre jeden sektor. Na presnejšiu analýzu by som pridal podmienku na počet sektorov v ktorých daná firma uverejnila ponuky.

Taktiež pri práci s `Vectorizer` by som samostatne spracoval nástroj R, pretože pri vektorizácii sa samostatné písmená odstránia.

Námet na pokračovanie by mohol spočívať v skúmaní podobných ponúk na Slovensku. Z rôznych pracovných portálov sa totiž dajú získať dáta podobné tým v mojej analýze. Následne by sme mohli výsledky porovnať.