Out[42]:

# Overview

## Dataset info

| | |
|---|---|
| **Number of variables** | 27 |
| **Number of observations** | 14586035 |
| **Total Missing (%)** | 0.4% |
| **Total size in memory** | 2.9 GiB |
| **Average record size in memory** | 216.0 B |

## Variables types

| | |
|---|---|
| **Numeric** | 13 |
| **Categorical** | 11 |
| **Boolean** | 2 |
| **Date** | 0 |
| **Text (Unique)** | 0 |
| **Rejected** | 1 |
| **Unsupported** | 0 |

## Warnings

- Cumulative_localclicks has 192822 / 1.3% zeros `Zeros`
- NAICS_Category has 1512773 / 10.4% missing values `Missing`
- avgOverallRating has 7810322 / 53.5% zeros `Zeros`
- city has a high cardinality: 17748 distinct values `Warning`
- clicks is highly skewed ($\gamma 1 = 43.392$) `Skewed`
- clicks has 214887 / 1.5% zeros `Zeros`
- companyId has a high cardinality: 150993 distinct values `Warning`
- date has a high cardinality: 395 distinct values `Warning`
- descriptionWordCount is highly correlated with descriptionCharacterLength ($\rho = 0.98859$) `Rejected`
- estimatedSalary has 257913 / 1.8% zeros `Zeros`
- experienceRequired has 9540289 / 65.4% zeros `Zeros`
- jobAgeDays has 406572 / 2.8% zeros `Zeros`
- jobDemand_Clicks is highly skewed ($\gamma 1 = 22.223$) `Skewed`
- jobDemand_localClicks has 192822 / 1.3% zeros `Zeros`
- jobId has a high cardinality: 520434 distinct values `Warning`
- localClicks is highly skewed ($\gamma 1 = 21.662$) `Skewed`
- localClicks has 7197480 / 49.3% zeros `Zeros`
- normTitle has a high cardinality: 6392 distinct values `Warning`
- normTitleCategory has a high cardinality: 58 distinct values `Warning`
- numReviews has 8248086 / 56.5% zeros `Zeros`
- stateProvince has a high cardinality: 87 distinct values `Warning`

# Variables

## Cumulative_clicks
Numeric

| | |
|---|---|
| **Distinct count** | 14377 |
| **Unique (%)** | 0.1% |
| Missing (%) | 0.0% |
| Missing (n) | 0 |
| Infinite (%) | 0.0% |
| Infinite (n) | 0 |
| **Mean** | 702.35 |
| **Minimum** | 2 |
| **Maximum** | 73460 |
| Zeros (%) | 0.0% |

Toggle details

## Cumulative_localclicks
Numeric

| | |
|---|---|
| **Distinct count** | 6132 |
| **Unique (%)** | 0.0% |
| Missing (%) | 0.0% |
| Missing (n) | 0 |
| Infinite (%) | 0.0% |
| Infinite (n) | 0 |
| **Mean** | 118.73 |
| **Minimum** | 0 |
| **Maximum** | 13399 |
| **Zeros (%)** | 1.3% |

Toggle details

## NAICS_Category
Categorical

| | |
|---|---|
| **Distinct count** | 18 |
| **Unique (%)** | 0.0% |
| **Missing (%)** | 10.4% |
| **Missing (n)** | 1512773 |

Professional, Scientific, and Technical S…        3007539

Retail Trade        1857512

Health Care and Social Assistance        1794076

Other values (14)                6414135

(Missing)        1512773

Toggle details

## Unnamed: 0
Numeric

| | |
|---|---|
| **Distinct count** | 14586035 |
| **Unique (%)** | 100.0% |
| Missing (%) | 0.0% |
| Missing (n) | 0 |
| Infinite (%) | 0.0% |
| Infinite (n) | 0 |
| **Mean** | 7293000 |
| **Minimum** | 1 |
| **Maximum** | 14586035 |
| Zeros (%) | 0.0% |

1e7

Toggle details

## avgOverallRating
Numeric

| | |
|---|---|
| **Distinct count** | 42 |
| **Unique (%)** | 0.0% |
| Missing (%) | 0.0% |
| Missing (n) | 0 |
| Infinite (%) | 0.0% |
| Infinite (n) | 0 |
| **Mean** | 1.7144 |
| **Minimum** | 0 |
| **Maximum** | 5 |
| **Zeros (%)** | 53.5% |
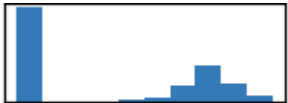
Toggle details

## city
Categorical

| | |
|---|---|
| **Distinct count** | 17748 |
| **Unique (%)** | 0.1% |
| Missing (%) | 0.0% |
| Missing (n) | 0 |

| | |
|---|---|
| Unknown | 338179 |
| New York | 229599 |
| Berlin | 140563 |
| Other values (17745) | 13877694 |

Toggle details

## clicks
Numeric

| | |
|---|---|
| **Distinct count** | 1827 |
| **Unique (%)** | 0.0% |
| Missing (%) | 0.0% |
| Missing (n) | 0 |
| Infinite (%) | 0.0% |
| Infinite (n) | 0 |
| **Mean** | 22.445 |
| **Minimum** | 0 |
| **Maximum** | 17196 |
| **Zeros (%)** | 1.5% |
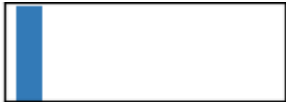
Toggle details

## companyId
Categorical

| | |
|---|---|
| **Distinct count** | 150993 |
| **Unique (%)** | 1.0% |
| Missing (%) | 0.0% |
| Missing (n) | 0 |

| | |
|---|---|
| company02436 | 222955 |
| company01921 | 70719 |
| company02209 | 59308 |
| Other values (150990) | 14233053 |

Toggle details

## country
Categorical

| | |
|---|---|
| **Distinct count** | 3 |
| **Unique (%)** | 0.0% |
| Missing (%) | 0.0% |
| Missing (n) | 0 |

| | |
|---|---|
| US | 11547904 |
| DE | 1925795 |
| CA | 1112336 |

Toggle details

## date
Categorical

| | |
|---|---|
| **Distinct count** | 395 |
| **Unique (%)** | 0.0% |
| Missing (%) | 0.0% |
| Missing (n) | 0 |

| | |
|---|---|
| 2017-08-24 | 42103 |
| 2017-08-25 | 42088 |
| 2017-08-23 | 41995 |
| Other values (392) | 14459849 |

Toggle details

## descriptionCharacterLength
Numeric

| | |
|---|---|
| **Distinct count** | 11075 |
| **Unique (%)** | 0.1% |
| Missing (%) | 0.0% |
| Missing (n) | 0 |
| Infinite (%) | 0.0% |
| Infinite (n) | 0 |
| **Mean** | 2367.6 |
| **Minimum** | 11 |
| **Maximum** | 57774 |
| Zeros (%) | 0.0% |

Toggle details

~~descriptionWordCount~~
Highly correlated

*This variable is highly correlated with <u>descriptionCharacterLength</u> and should be ignored for analysis*

**Correlation**      0.98859

---

## educationRequirements
Categorical

**Distinct count**      3
**Unique (%)**          0.0%
Missing (%)             0.0%
Missing (n)             0

|                  |         |
|------------------|---------|
| None             | 8428772 |
| Higher Education | 3282252 |
| High School      | 2875011 |

Toggle details

---

## estimatedSalary
Numeric

**Distinct count**      2060
**Unique (%)**          0.0%
Missing (%)             0.0%
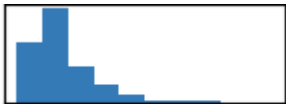Missing (n)             0
Infinite (%)            0.0%
Infinite (n)            0
**Mean**                43669
**Minimum**             0
**Maximum**             248900
**Zeros (%)**           1.8%



Toggle details

---

## experienceRequired
Numeric

**Distinct count**      42
**Unique (%)**          0.0%
Missing (%)             0.0%
Missing (n)             0
Infinite (%)            0.0%

| | |
|---|---|
| Infinite (n) | 0 |
| **Mean** | 1.1751 |
| **Minimum** | 0 |
| **Maximum** | 20 |
| **Zeros (%)** | 65.4% |



Toggle details

## jobAgeDays
Numeric

| | |
|---|---|
| **Distinct count** | 102 |
| **Unique (%)** | 0.0% |
| Missing (%) | 0.0% |
| Missing (n) | 0 |
| Infinite (%) | 0.0% |
| Infinite (n) | 0 |
| **Mean** | 30.777 |
| **Minimum** | 0 |
| **Maximum** | 101 |
| **Zeros (%)** | 2.8% |



Toggle details

## jobDemand_Clicks
Numeric

| | |
|---|---|
| **Distinct count** | 39737 |
| **Unique (%)** | 0.3% |
| Missing (%) | 0.0% |
| Missing (n) | 0 |
| Infinite (%) | 0.0% |
| Infinite (n) | 0 |
| **Mean** | 29.294 |
| **Minimum** | 0.02 |
| **Maximum** | 17165 |
| Zeros (%) | 0.0% |



Toggle details

## jobDemand_localClicks
Numeric

| | |
|---|---|
| **Distinct count** | 17392 |
| **Unique (%)** | 0.1% |
| Missing (%) | 0.0% |
| Missing (n) | 0 |
| Infinite (%) | 0.0% |
| Infinite (n) | 0 |
| **Mean** | 5.2112 |
| **Minimum** | 0 |
| **Maximum** | 2048 |
| **Zeros (%)** | 1.3% |

Toggle details

## jobId
Categorical

| | |
|---|---|
| **Distinct count** | 520434 |
| **Unique (%)** | 3.6% |
| Missing (%) | 0.0% |
| Missing (n) | 0 |

| | |
|---|---|
| job0210510 | 102 |
| job0246932 | 102 |
| job0165719 | 101 |
| Other values (520431) | 14585730 |

Toggle details

## jobLanguage
Categorical

| | |
|---|---|
| **Distinct count** | 3 |
| **Unique (%)** | 0.0% |
| Missing (%) | 0.0% |
| Missing (n) | 0 |

| | |
|---|---|
| EN | 12547718 |
| DE | 1853422 |
| FR | 184895 |

Toggle details

## licenseRequiredJob
Boolean

| | |
|---|---|
| **Distinct count** | 2 |
| **Unique (%)** | 0.0% |
| Missing (%) | 0.0% |
| Missing (n) | 0 |
| **Mean** | 0.30772 |

| | |
|---|---|
| 0.0 | 10097668 |
| 1.0 | 4488367 |

Toggle details

## localClicks
Numeric

| | |
|---|---|
| **Distinct count** | 764 |
| **Unique (%)** | 0.0% |
| Missing (%) | 0.0% |
| Missing (n) | 0 |
| Infinite (%) | 0.0% |
| Infinite (n) | 0 |
| **Mean** | 3.7098 |
| **Minimum** | 0 |
| **Maximum** | 2156 |
| **Zeros (%)** | 49.3% |

Toggle details

## normTitle
Categorical

| | |
|---|---|
| **Distinct count** | 6392 |
| **Unique (%)** | 0.0% |
| Missing (%) | 0.0% |
| Missing (n) | 0 |

| | |
|---|---|
| Unknown | 1302545 |
| retail sales associate | 291641 |
| customer service representative | 196005 |
| Other values (6389) | 12795844 |

Toggle details

## normTitleCategory
Categorical

| | |
|---|---|
| **Distinct count** | 58 |
| **Unique (%)** | 0.0% |
| Missing (%) | 0.0% |
| Missing (n) | 0 |

| | |
|---|---|
| Unknown | 1302545 |
| management | 1103743 |
| retail | 931695 |
| Other values (55) | 11248052 |

Toggle details

## numReviews
Numeric

| | |
|---|---|
| **Distinct count** | 631 |
| **Unique (%)** | 0.0% |
| Missing (%) | 0.0% |
| Missing (n) | 0 |
| Infinite (%) | 0.0% |
| Infinite (n) | 0 |
| **Mean** | 81.379 |
| **Minimum** | 0 |
| **Maximum** | 10565 |
| **Zeros (%)** | 56.5% |

Toggle details

## stateProvince
Categorical

| | |
|---|---|
| **Distinct count** | 87 |
| **Unique (%)** | 0.0% |
| Missing (%) | 0.0% |
| Missing (n) | 0 |

| | |
|---|---|
| CA | 1336804 |
| TX | 906323 |
| FL | 688840 |
| Other values (84) | 11654068 |

Toggle details

## supervisingJob
Boolean

| | |
|---|---|
| **Distinct count** | 2 |
| **Unique (%)** | 0.0% |
| Missing (%) | 0.0% |
| Missing (n) | 0 |
| **Mean** | 0.13492 |

| | |
|---|---|
| 0.0 | 12618103 |
| 1.0 | 1967932 |

Toggle details

# Correlations


Pearson

## Spearman
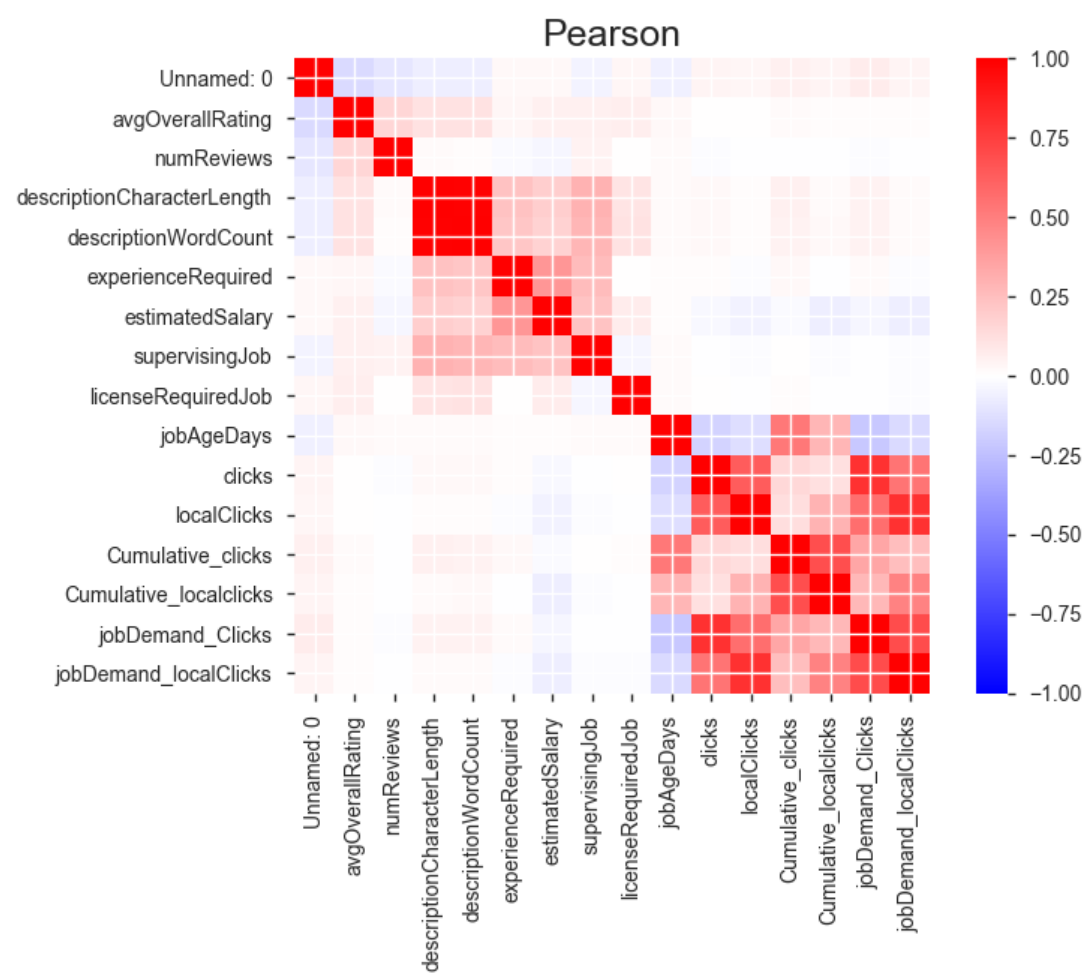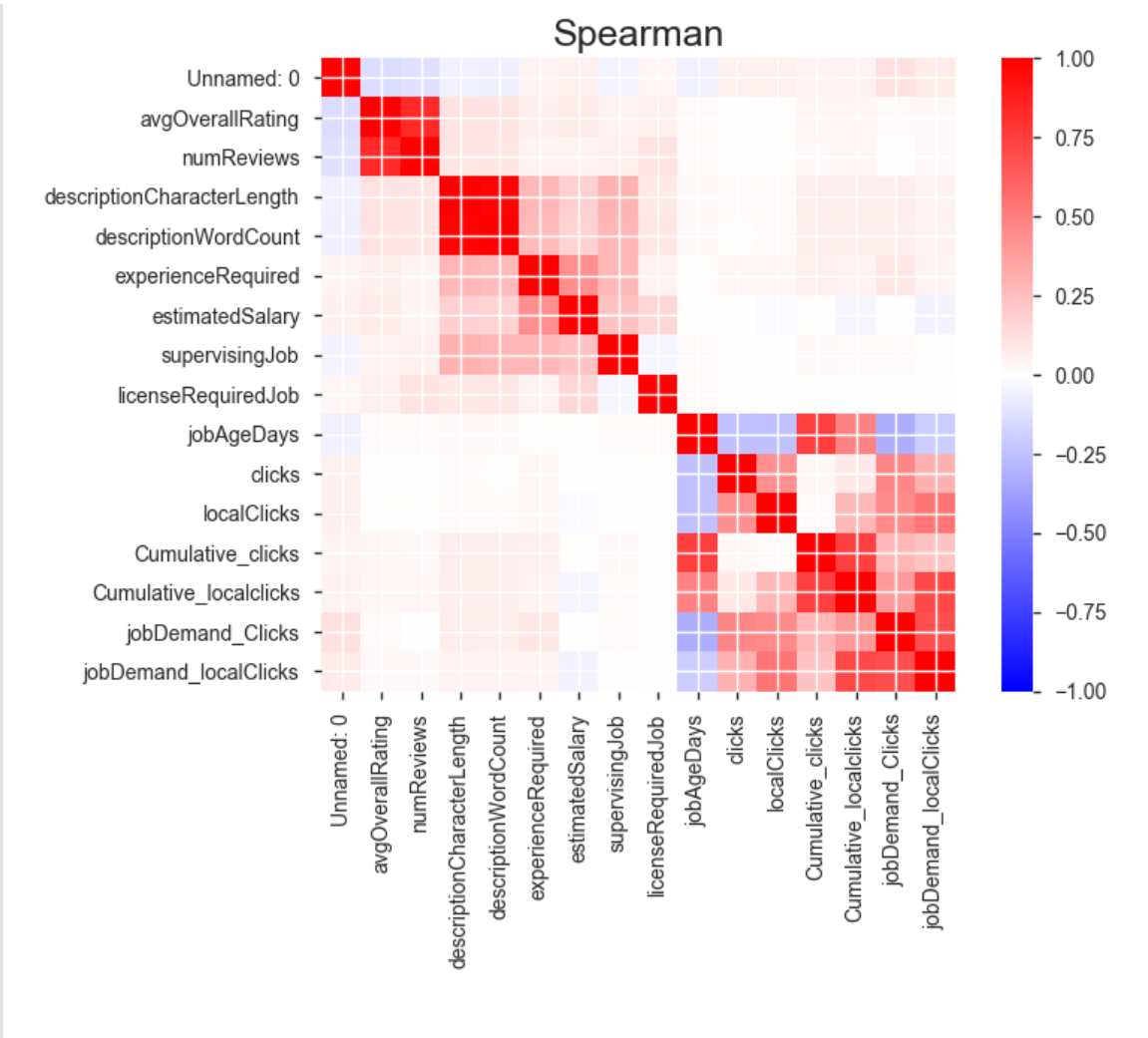


# Sample

|   | Unnamed: 0 | date | companyId | jobId | country | stateP... |
|---|---|---|---|---|---|---|
| **0** | 1 | 2016-11-01 | company00000 | job0000000 | CA | ON |
| **1** | 2 | 2016-11-02 | company00001 | job0016247 | US | PA |
| **2** | 3 | 2016-11-03 | company00001 | job0016247 | US | PA |
| **3** | 4 | 2016-11-04 | company00001 | job0016247 | US | PA |
| **4** | 5 | 2016-11-05 | company00001 | job0016247 | US | PA |