In [39]:
```
import pandas as pd
```

In [40]:
```
#Loading dataset
df=pd.read_csv('data0/aug_train.csv')
df.describe()
```

Out[40]:

	enrollee_id	city_development_index	training_hours	target
count	19158.000000	19158.000000	19158.000000	19158.000000
mean	16875.358179	0.828848	65.366896	0.249348
std	9616.292592	0.123362	60.058462	0.432647
min	1.000000	0.448000	1.000000	0.000000
25%	8554.250000	0.740000	23.000000	0.000000
50%	16982.500000	0.903000	47.000000	0.000000
75%	25169.750000	0.920000	88.000000	0.000000
max	33380.000000	0.949000	336.000000	1.000000

In [41]:
```
df['city'].describe()
```

Out[41]:
```
count        19158
unique         123
top       city_103
freq          4355
Name: city, dtype: object
```

EDA Report with Pandas Profiling

In [42]:
```
import pandas_profiling
```

In [43]:
```
profile_report=df.profile_report()
```

In [44]:
```
profile_report
```

```
Summarize dataset:   0%|          | 0/5 [00:00<?, ?it/s]
Generate report structure:   0%|          | 0/1 [00:00<?, ?it/s]
```

Render HTML: 0%| | 0/1 [00:00<?, ?it/s]

Overview

Dataset statistics

Number of variables	14
Number of observations	19158
Missing cells	20733
Missing cells (%)	7.7%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	2.0 MiB
Average record size in memory	112.0 B

Variable types

Numeric	3
Categorical	11

Alerts

`city` has a high cardinality: 123 distinct values	**High cardinality**
`relevent_experience` is highly correlated with `experience` and 1 other fields (experience, last_new_job)	**High correlation**
`experience` is highly correlated with `relevent_experience`	**High correlation**
`last_new_job` is highly correlated with `relevent_experience`	**High correlation**
`gender` has 4508 (23.5%) missing values	**Missing**

enrolled_university	has 386 (2.0%) missing values

Missing

Out[44]:

In [45]:
```
profile_report.to_file('data0.pdf')
```

```
Export report to file:   0%|          | 0/1 [00:00<?, ?it/s]
```

EDA Report with Sweetviz

In [46]:
```
import sweetviz as sv
```

In [47]:
```
sweetviz_report=sv.analyze(df)
```

```
                                             |                                             | [  0%]   00:00 ->…
```

In [48]:
```
sweetviz_report.show_html()
```

```
Report SWEETVIZ_REPORT.html was generated! NOTEBOOK/COLAB USERS: the web browser MAY not pop up, regardless, the report
IS saved in your notebook/colab files.
```

EDA Report with Autotviz

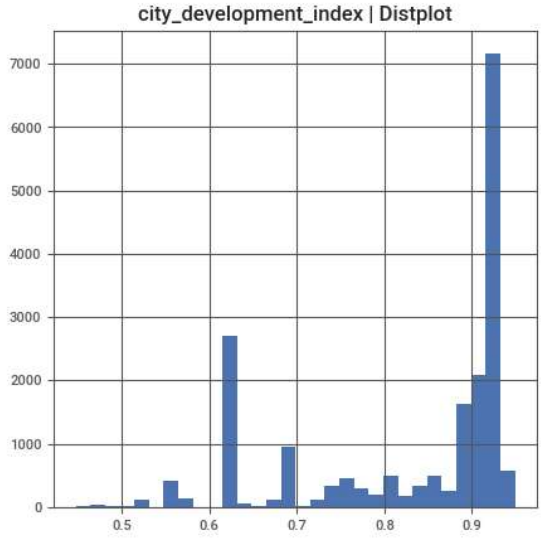
In [49]:
```
from autoviz.AutoViz_Class import AutoViz_Class
```

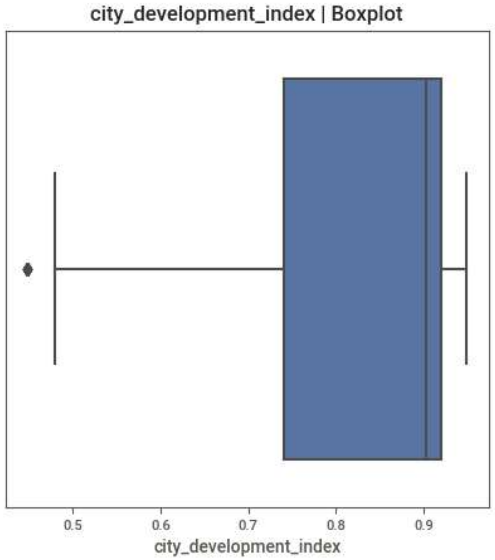
In [50]:
```
#EDA using Autoviz
av=AutoViz_Class()
data = av.AutoViz('data0/aug_train.csv')
```

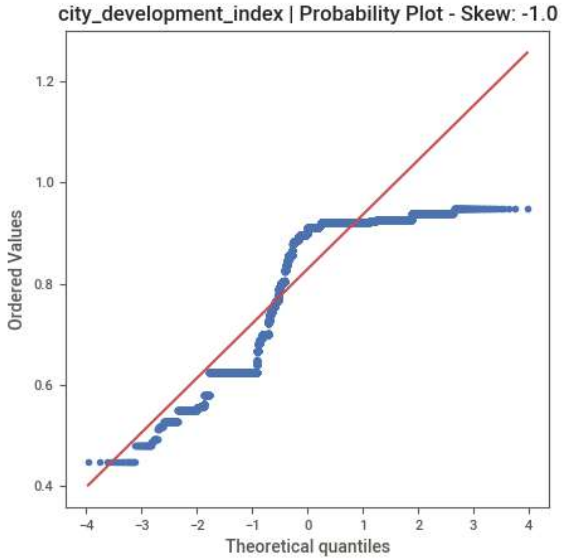
```
Shape of your Data Set loaded: (19158, 14)
############################################################################################
######################## C L A S S I F Y I N G  V A R I A B L E S  ##########################
############################################################################################
Classifying variables in data set...
```

	Nuniques	dtype	Nulls	Nullpercent	NuniquePercent	Value counts Min	Data cleaning improvement suggestions
enrollee_id	19158	int64	0	0.000000	100.000000	0	possible ID column: drop
training_hours	241	int64	0	0.000000	1.257960	0	
city	123	object	0	0.000000	0.642029	1	combine rare categories
city_development_index	93	float64	0	0.000000	0.485437	0	
experience	22	object	65	0.339284	0.114835	148	fill missing values, fix mixed data types
company_size	8	object	5938	30.994885	0.041758	563	fill missing values, fix mixed data types
major_discipline	6	object	2813	14.683161	0.031319	223	fill missing values, fix mixed data types
company_type	6	object	6140	32.049274	0.031319	121	fill missing values, fix mixed data types
last_new_job	6	object	423	2.207955	0.031319	1024	fill missing values, fix mixed data types
education_level	5	object	460	2.401086	0.026099	308	fill missing values, fix mixed data types
gender	3	object	4508	23.530640	0.015659	191	fill missing values, fix mixed data types
enrolled_university	3	object	386	2.014824	0.015659	1198	fill missing values, fix mixed data types
relevent_experience	2	object	0	0.000000	0.010440	5366	
target	2	float64	0	0.000000	0.010440	0	skewed column: cap or drop possible outliers

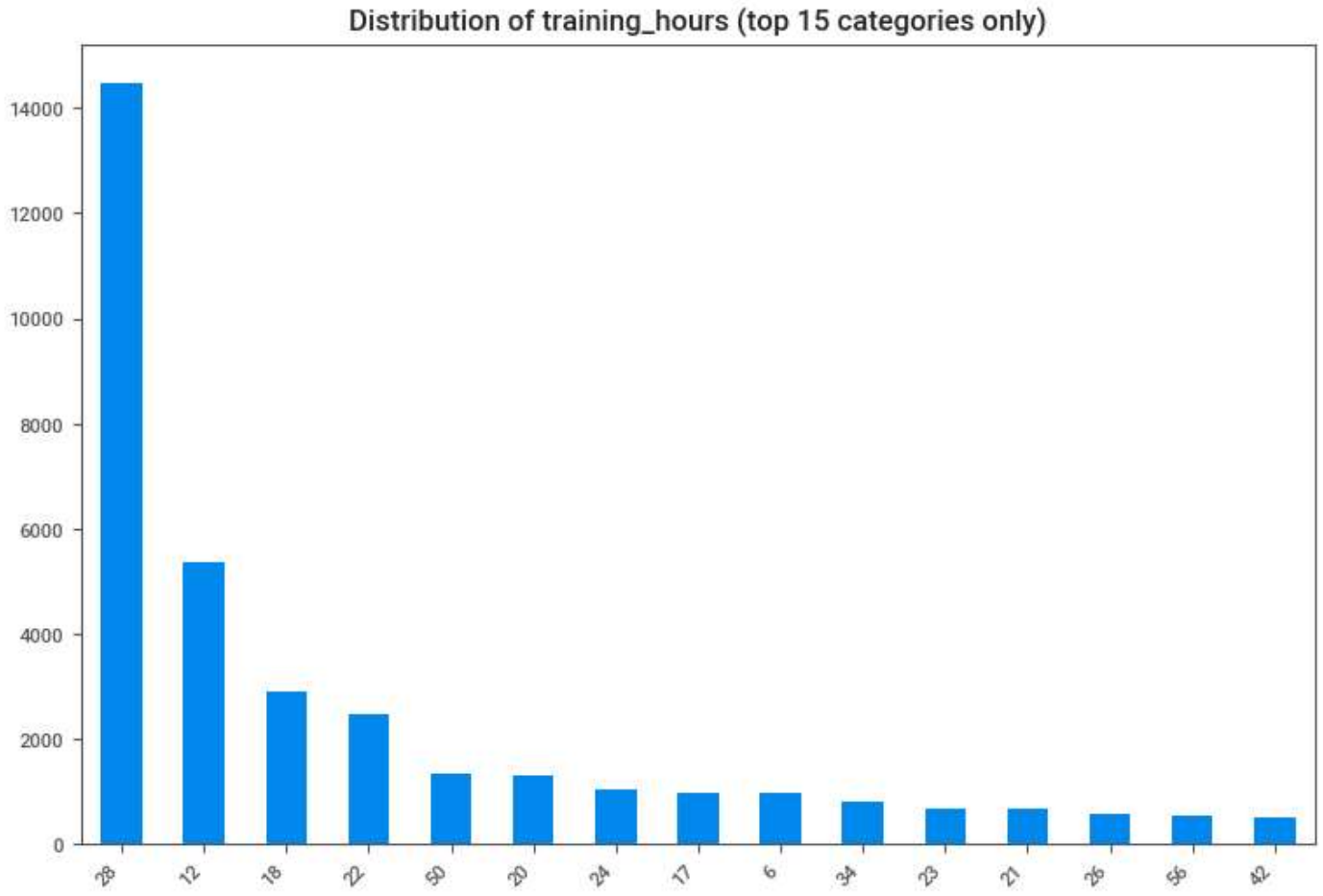
```
14 Predictors classified...
    1 variables removed since they were ID or low-information variables
    List of variables removed: ['enrollee_id']
```


city_development_index | Distplot
7000
6000
5000
4000
3000
2000
1000
0
0.5
0.6
0.7
0.8
0.9

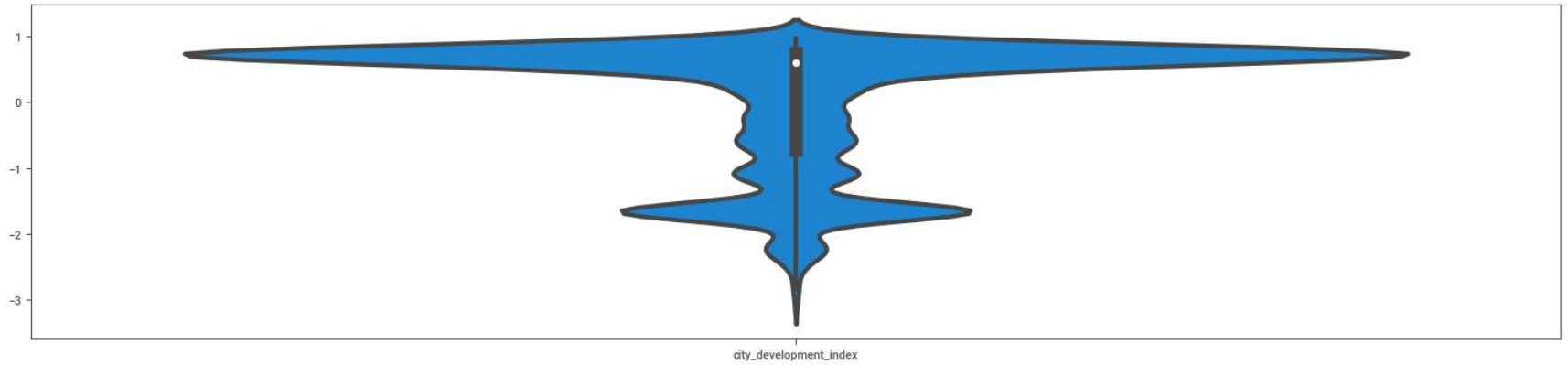

city_development_index | Boxplot
0.5
0.6
0.7
0.8
0.9
city_development_index

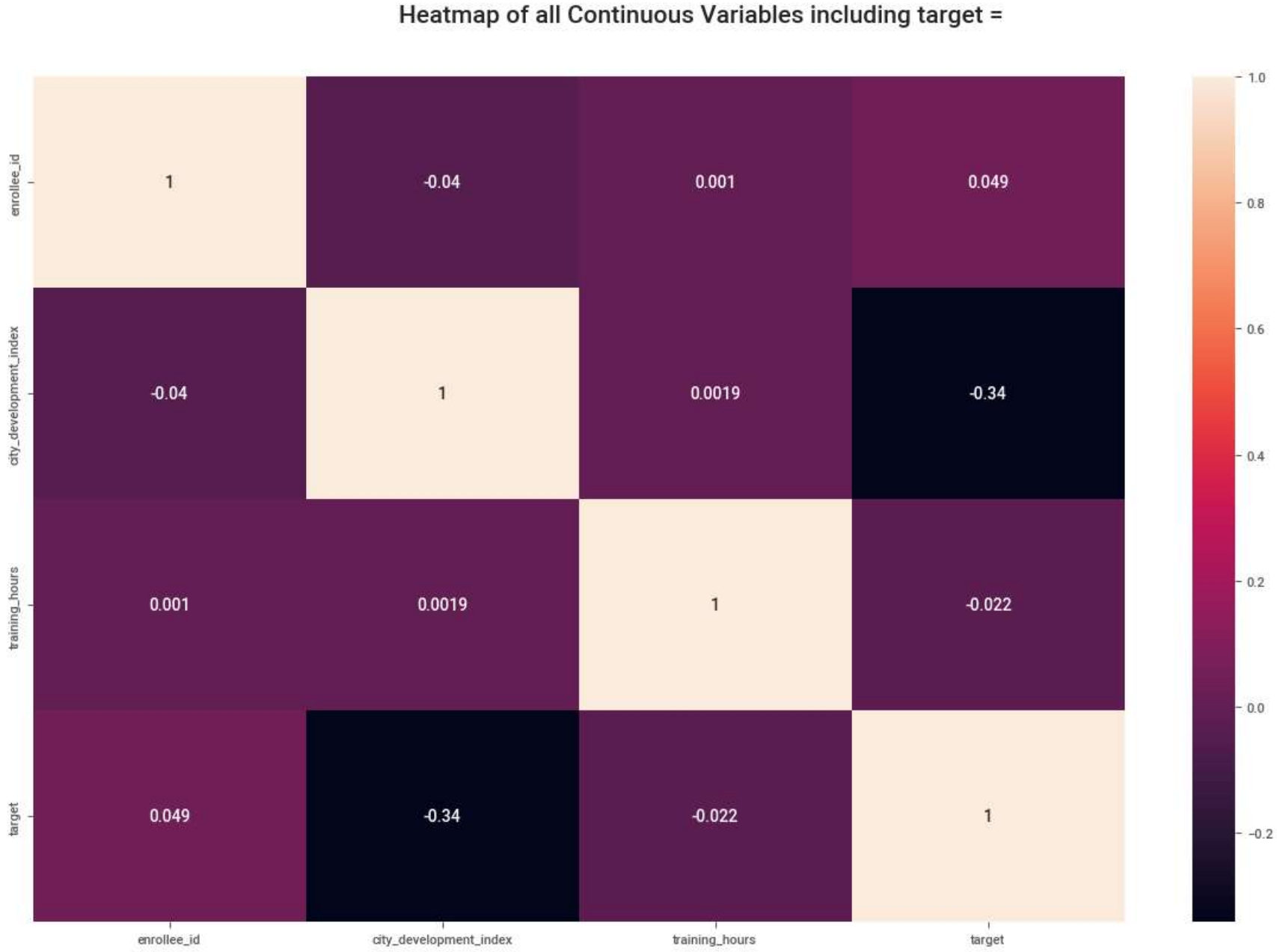

city_development_index | Probability Plot - Skew: -1.0
Ordered Values
1.2
1.0
0.8
0.6
0.4
−4
−3
−2
−1
0
1
2
3
4
Theoretical quantiles

Histograms (KDE plots) of all Continuous Variables

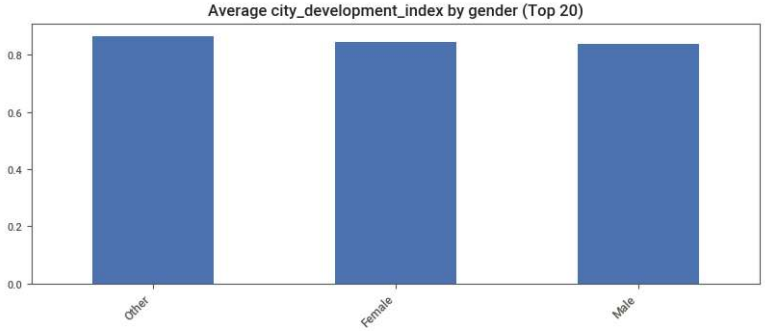


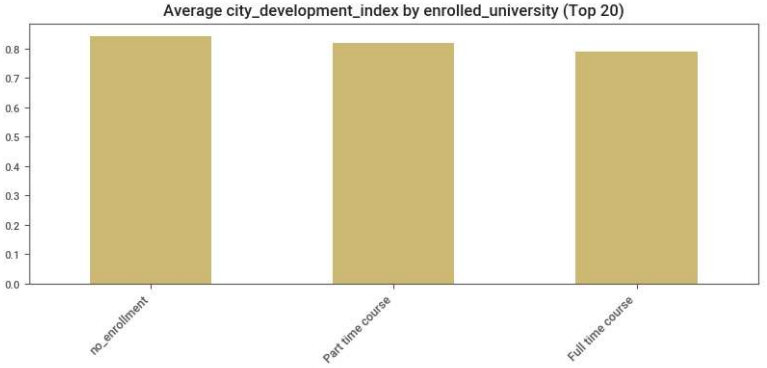
Violin Plot of all Continuous Variables

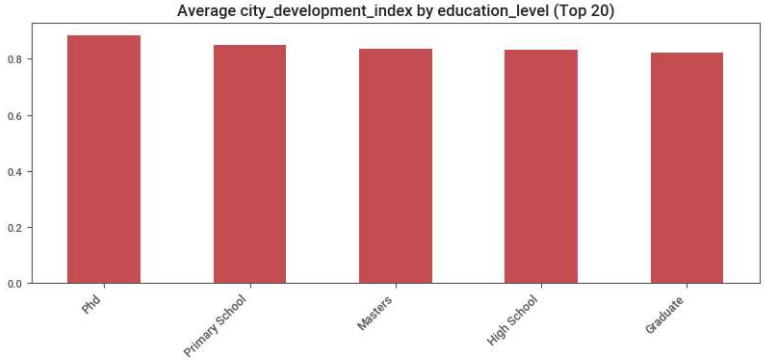


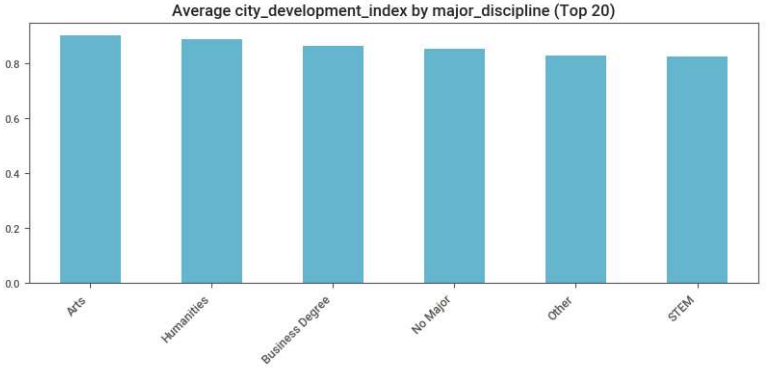

Heatmap of all Continuous Variables including target =
enrollee_id
city_development_index
training_hours
target
1
-0.04
0.001
0.049
-0.04
1
0.0019
-0.34
0.001
0.0019
1
-0.022
0.049
-0.34
-0.022
1
enrollee_id
city_development_index
training_hours
target
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

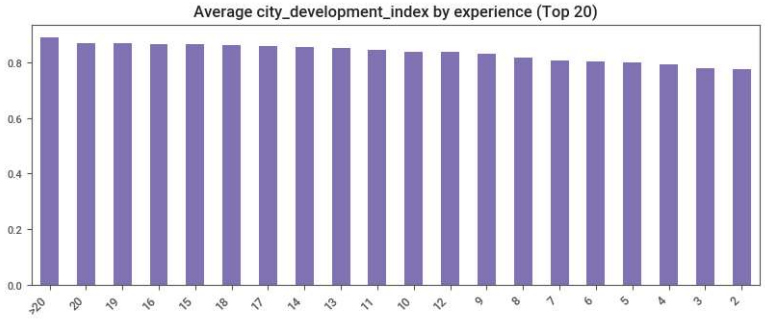
Bar plots for each Continuous by each Categorical variable

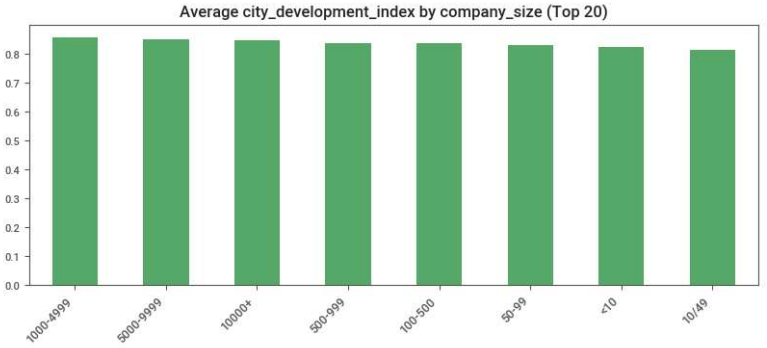






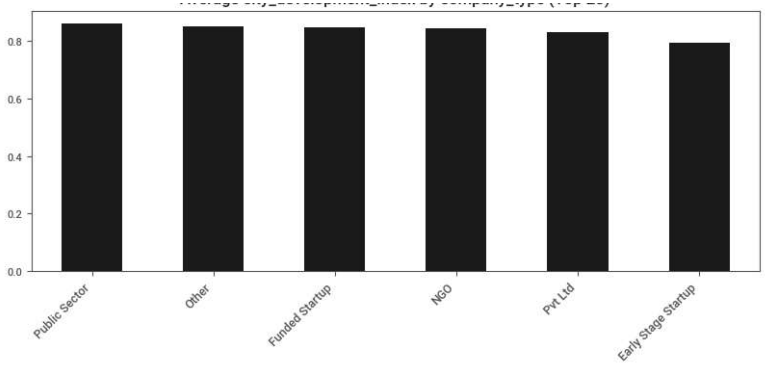


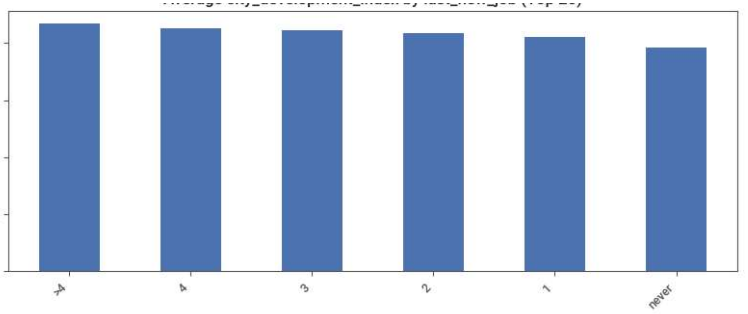




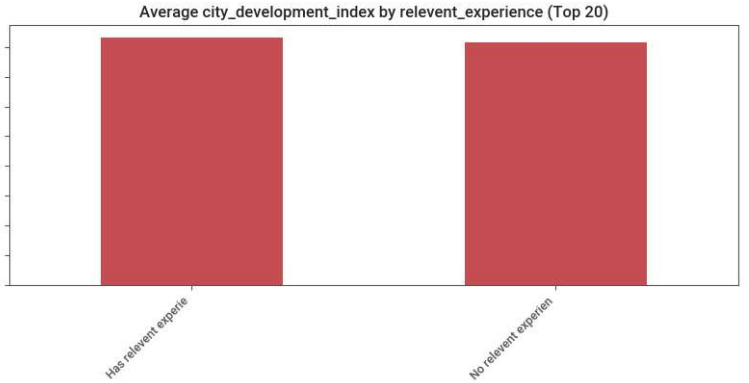
Average city_development_index by company_type (Top 20)

Average city_development_index by last_new_job (Top 20)


0.0
0.2
0.4
0.6
0.8
Public Sector
Other
Funded Startup
NGO
Pvt Ltd
Early Stage Startup

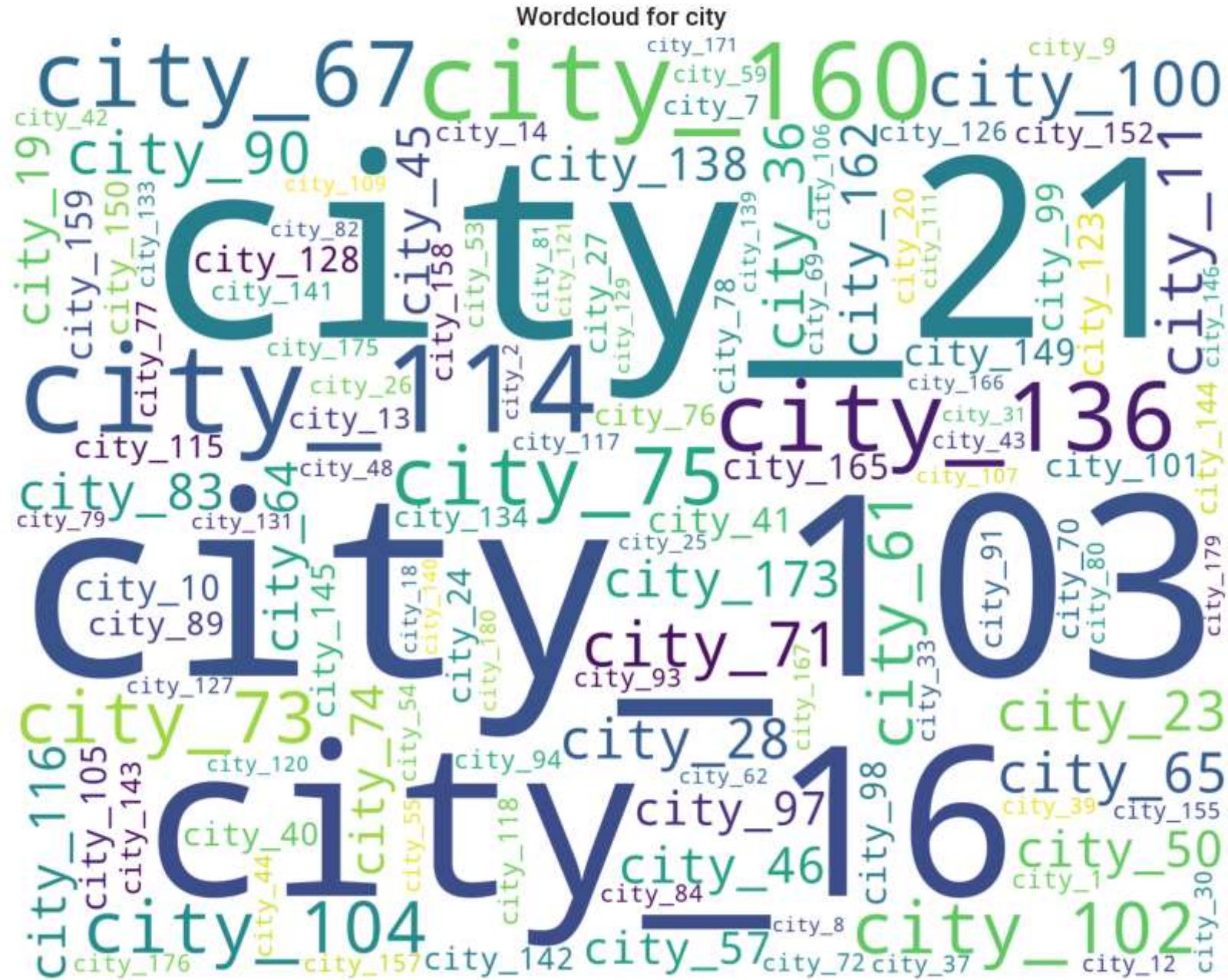

0.0
0.2
0.4
0.6
0.8
>4
4
3
2
1
never


Average city_development_index by training_hours (Top 20)
0.0
0.2
0.4
0.6
0.8
276
262
322
312
120
226
212
125
314
278
206
256
152
178
127
130
336
200
90
260


Average city_development_index by relevent_experience (Top 20)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Has relevent experie
No relevent experien

```
[nltk_data] Downloading collection 'popular'
[nltk_data]    |
[nltk_data]    | Downloading package cmudict to
[nltk_data]    |     C:\Users\dpkjs\AppData\Roaming\nltk_data...
[nltk_data]    |   Package cmudict is already up-to-date!
[nltk_data]    | Downloading package gazetteers to
[nltk_data]    |     C:\Users\dpkjs\AppData\Roaming\nltk_data...
[nltk_data]    |   Package gazetteers is already up-to-date!
[nltk_data]    | Downloading package genesis to
[nltk_data]    |     C:\Users\dpkjs\AppData\Roaming\nltk_data...
[nltk_data]    |   Package genesis is already up-to-date!
[nltk_data]    | Downloading package gutenberg to
[nltk_data]    |     C:\Users\dpkjs\AppData\Roaming\nltk_data...
[nltk_data]    |   Package gutenberg is already up-to-date!
[nltk_data]    | Downloading package inaugural to
[nltk_data]    |     C:\Users\dpkjs\AppData\Roaming\nltk_data...
[nltk_data]    |   Package inaugural is already up-to-date!
[nltk_data]    | Downloading package movie_reviews to
[nltk_data]    |     C:\Users\dpkjs\AppData\Roaming\nltk_data...
[nltk_data]    |   Package movie_reviews is already up-to-date!
[nltk_data]    | Downloading package names to
[nltk_data]    |     C:\Users\dpkjs\AppData\Roaming\nltk_data...
[nltk_data]    |   Package names is already up-to-date!
[nltk_data]    | Downloading package shakespeare to
[nltk_data]    |     C:\Users\dpkjs\AppData\Roaming\nltk_data...
[nltk_data]    |   Package shakespeare is already up-to-date!
[nltk_data]    | Downloading package stopwords to
[nltk_data]    |     C:\Users\dpkjs\AppData\Roaming\nltk_data...
[nltk_data]    |   Package stopwords is already up-to-date!
[nltk_data]    | Downloading package treebank to
[nltk_data]    |     C:\Users\dpkjs\AppData\Roaming\nltk_data...
[nltk_data]    |   Package treebank is already up-to-date!
[nltk_data]    | Downloading package twitter_samples to
[nltk_data]    |     C:\Users\dpkjs\AppData\Roaming\nltk_data...
[nltk_data]    |   Package twitter_samples is already up-to-date!
[nltk_data]    | Downloading package omw to
[nltk_data]    |     C:\Users\dpkjs\AppData\Roaming\nltk_data...
[nltk_data]    |   Package omw is already up-to-date!
[nltk_data]    | Downloading package omw-1.4 to
[nltk_data]    |     C:\Users\dpkjs\AppData\Roaming\nltk_data...
[nltk_data]    |   Package omw-1.4 is already up-to-date!
[nltk_data]    | Downloading package wordnet to
[nltk_data]    |     C:\Users\dpkjs\AppData\Roaming\nltk_data...
[nltk_data]    |   Package wordnet is already up-to-date!
[nltk_data]    | Downloading package wordnet2021 to
```

```
[nltk_data]    |     C:\Users\dpkjs\AppData\Roaming\nltk_data...
[nltk_data]    |   Package wordnet2021 is already up-to-date!
[nltk_data]    | Downloading package wordnet31 to
[nltk_data]    |     C:\Users\dpkjs\AppData\Roaming\nltk_data...
[nltk_data]    |   Package wordnet31 is already up-to-date!
[nltk_data]    | Downloading package wordnet_ic to
[nltk_data]    |     C:\Users\dpkjs\AppData\Roaming\nltk_data...
[nltk_data]    |   Package wordnet_ic is already up-to-date!
[nltk_data]    | Downloading package words to
[nltk_data]    |     C:\Users\dpkjs\AppData\Roaming\nltk_data...
[nltk_data]    |   Package words is already up-to-date!
[nltk_data]    | Downloading package maxent_ne_chunker to
[nltk_data]    |     C:\Users\dpkjs\AppData\Roaming\nltk_data...
[nltk_data]    |   Package maxent_ne_chunker is already up-to-date!
[nltk_data]    | Downloading package punkt to
[nltk_data]    |     C:\Users\dpkjs\AppData\Roaming\nltk_data...
[nltk_data]    |   Package punkt is already up-to-date!
[nltk_data]    | Downloading package snowball_data to
[nltk_data]    |     C:\Users\dpkjs\AppData\Roaming\nltk_data...
[nltk_data]    |   Package snowball_data is already up-to-date!
[nltk_data]    | Downloading package averaged_perceptron_tagger to
[nltk_data]    |     C:\Users\dpkjs\AppData\Roaming\nltk_data...
[nltk_data]    |   Package averaged_perceptron_tagger is already up-
[nltk_data]    |       to-date!
[nltk_data]    |
[nltk_data]  Done downloading collection popular
```



All Plots done
Time to run AutoViz = 11 seconds

###################### AUTO VISUALIZATION Completed ######################

EDA Report with Dtale

In [51]:
```
import dtale
```

In [52]:
```
dtale.show(df)
```

	0
0	

Out[52]:

EDA Report with DataPrep

In [53]:
```
from dataprep.eda import create_report
```

In [54]:
```
create_report(df)
```

```
  0%|                                                                                                | 0/1603 [00:00<…
```

Out[54]:

Overview

Dataset Statistics

Number of Variables	14
Number of Rows	19158
Missing Cells	20733
Missing Cells (%)	7.7%
Duplicate Rows	0
Duplicate Rows (%)	0.0%
Total Size in Memory	11.8 MB
Average Row Size in Memory	646.9 B
Variable Types	Numerical: 3 Categorical: 11

Variables

Sort by Feature order ☐ Reverse order

enrollee_id
numerical

Show Details

Approximate Distinct Count	19158
Approximate Unique (%)	100.0%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Memory Size	306528

Mean	16875.3582
Minimum	1
Maximum	33380
Zeros	0
Zeros (%)	0.0%
Negatives	0
Negatives (%)	0.0%

city
categorical

Show Details

Approximate Distinct Count	123
Approximate Unique (%)	0.6%
Missing	0
Missing (%)	0.0%
Memory Size	1389152

Approximate Distinct	93	Count

Approximate Unique (%)	0.5%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Memory Size	306528
Mean	0.8288

Minimum	0.448
Maximum	0.949
Zeros	0
Zeros (%)	0.0%
Negatives	0
Negatives (%)	0.0%

city_development_ind…
numerical

Show Details

Approximate Distinct Count	3
Approximate Unique (%)	0.0%
Missing	4508
Missing (%)	23.5%
Memory Size	1013517

gender
categorical

Show Details

Approximate Distinct Count	2

relevent_experience
categorical

Show Details

Approximate Unique (%)	0.0%
Missing	0
Missing (%)	0.0%
Memory Size	1680538

enrolled_university
categorical

Show Details

Approximate Distinct Count	3
Approximate Unique (%)	0.0%
Missing	386
Missing (%)	2.0%
Memory Size	1479081

education_level

Approximate Distinct Count	5
Approximate Unique (%)	0.0%

categorical

Show Details

Missing	460
Missing (%)	2.4%
Memory Size	1366422

major_discipline
categorical

Show Details

Approximate Distinct Count	6
Approximate Unique (%)	0.0%
Missing	2813
Missing (%)	14.7%
Memory Size	1136689

experience
categorical

Approximate Distinct Count	22
Approximate Unique (%)	0.1%
Missing	65

Show Details

Missing (%)	0.3%
Memory Size	1272628

company_size
categorical

Show Details

Approximate Distinct Count	8
Approximate Unique (%)	0.1%
Missing	5938
Missing (%)	31.0%
Memory Size	939263

company_type
categorical

Show Details

Approximate Distinct Count	6
Approximate Unique (%)	0.0%
Missing	6140
Missing (%)	32.0%

Memory Size	954943

last_new_job
categorical

Show Details

Approximate Distinct Count	6
Approximate Unique (%)	0.0%
Missing	423
Missing (%)	2.2%
Memory Size	1249608

training_hours
numerical

Show Details

Approximate Distinct Count	241
Approximate Unique (%)	1.3%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%

Mean	65.3669
Minimum	1
Maximum	336
Zeros	0
Zeros (%)	0.0%
Negatives	0

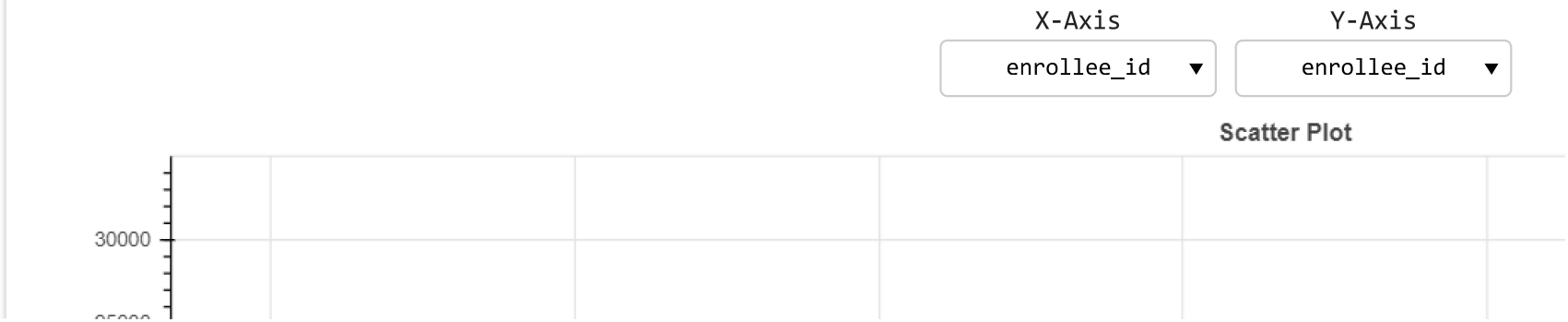
Memory Size	306528	Negatives (%)	0.0%

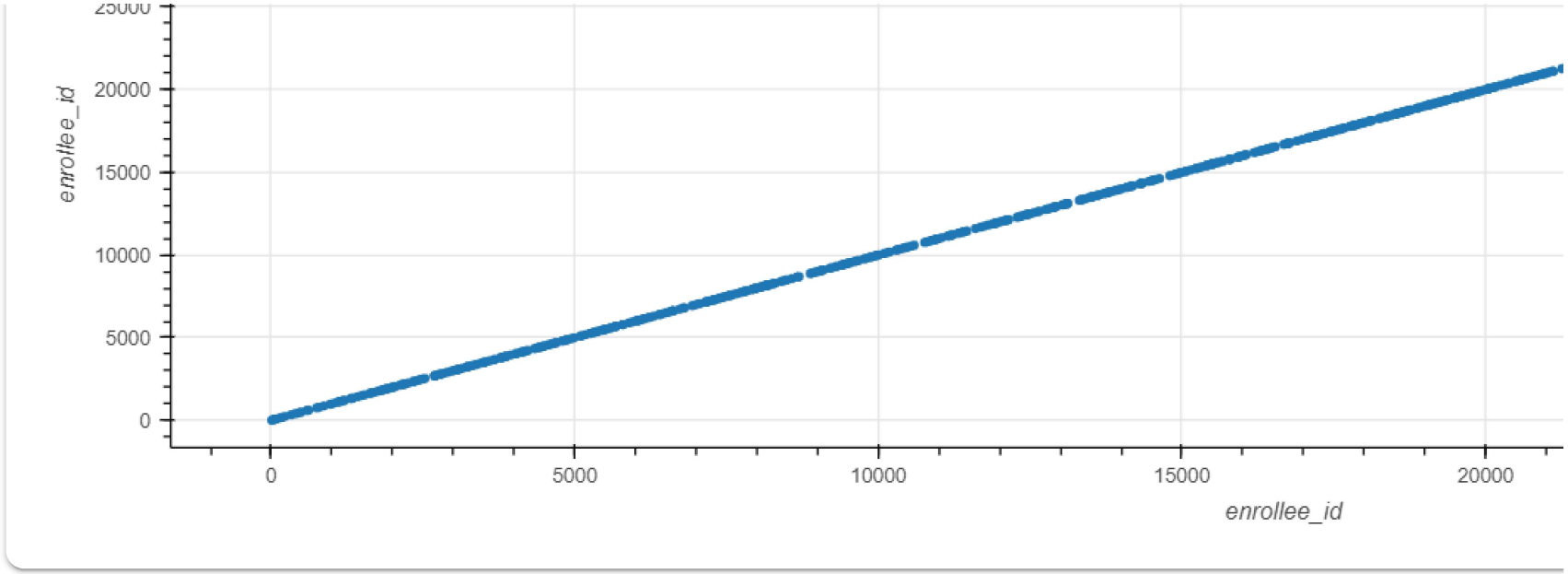
target
categorical

Show Details

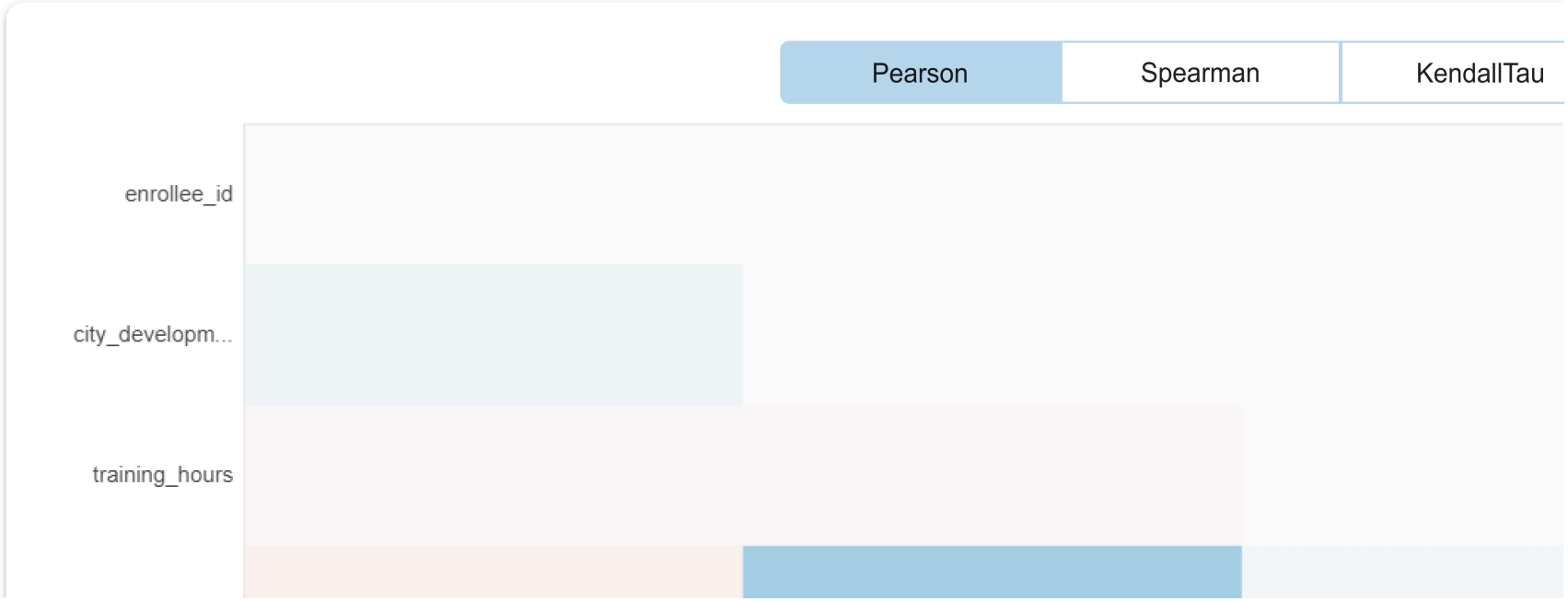
Approximate Distinct Count	2
Approximate Unique (%)	0.0%
Missing	0
Missing (%)	0.0%
Memory Size	1302744

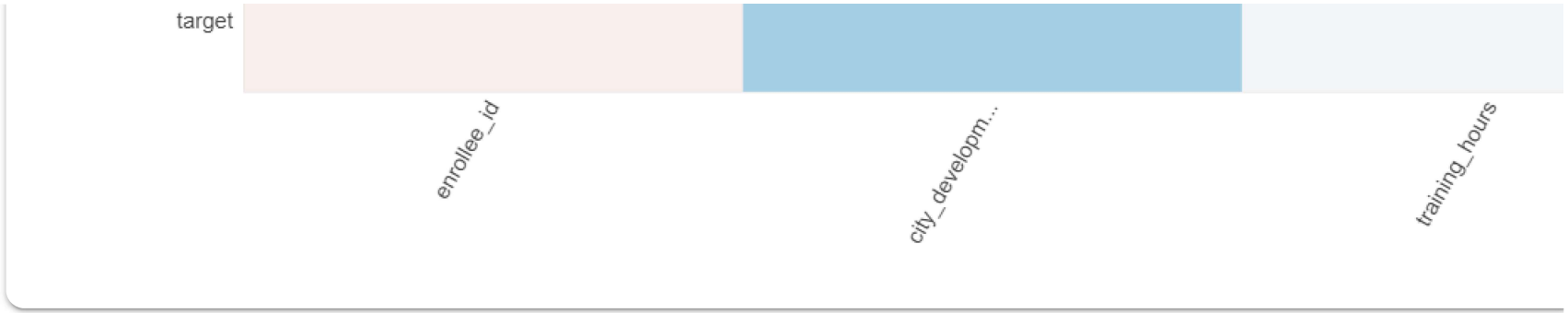
Interactions



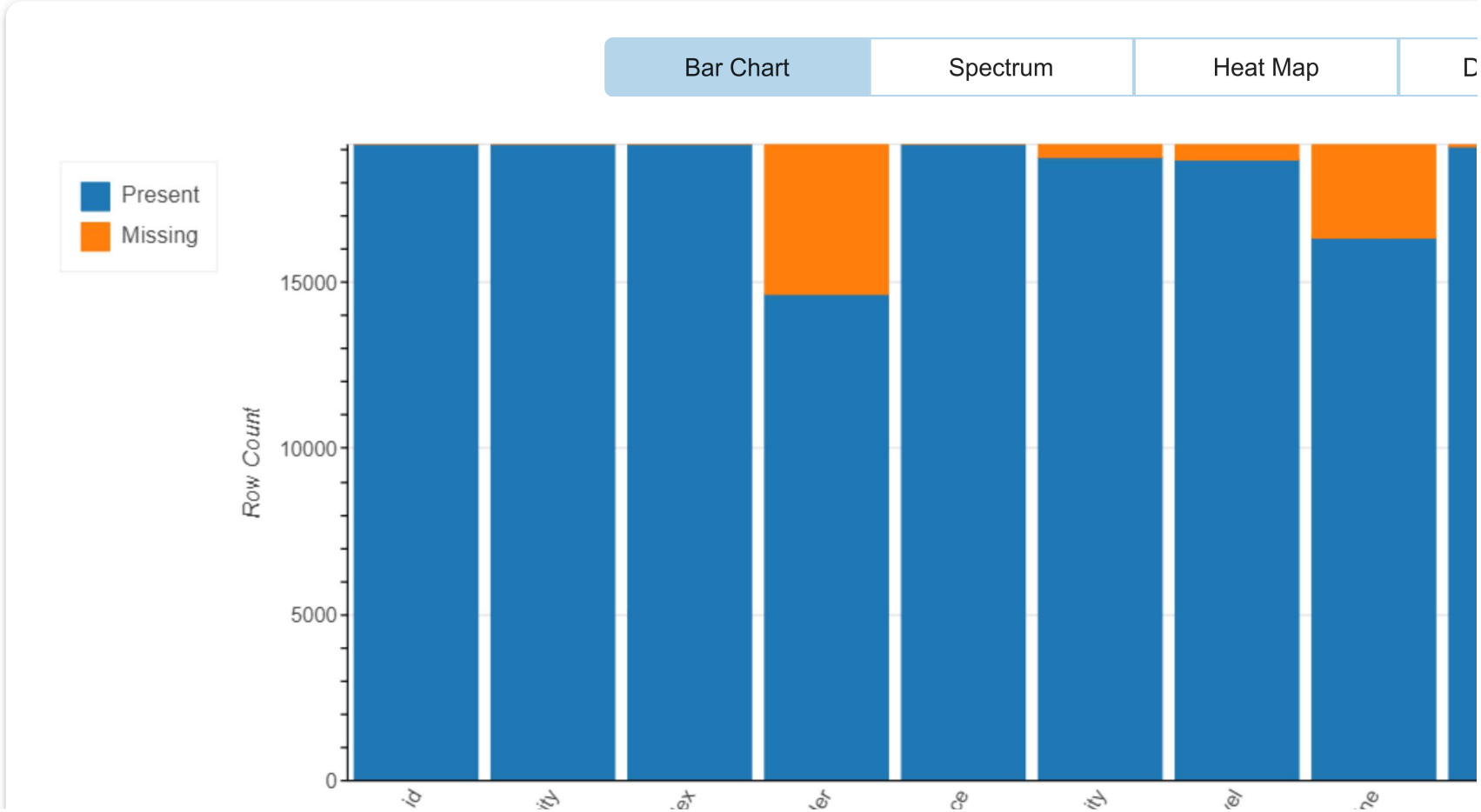


Correlations





Missing Values



Report generated with DataPrep

EDA Report with statsModels

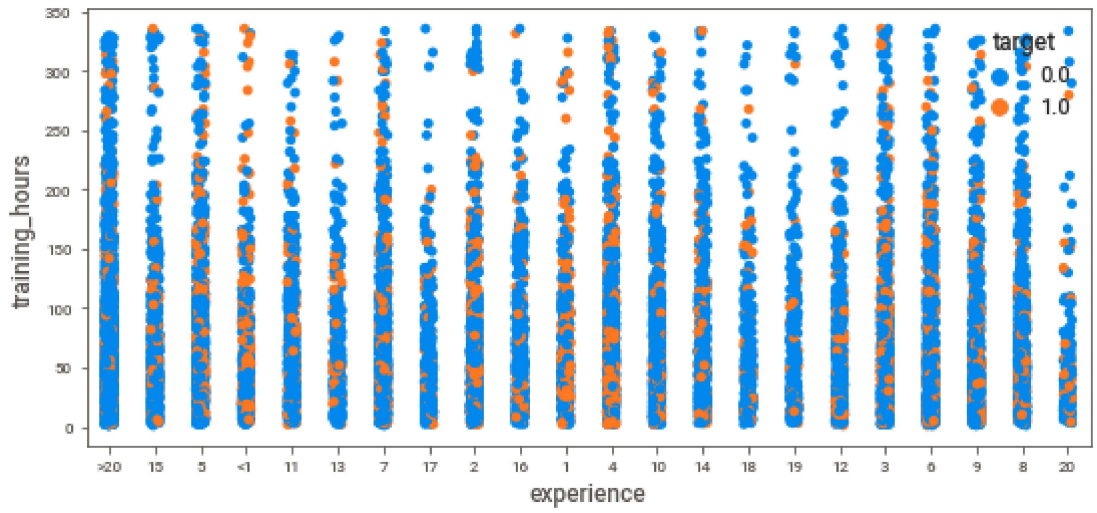
In [55]:
```
from speedml import Speedml
sml = Speedml('data0/aug_train.csv',
              'data0/aug_test.csv',
             target='target')

sml.eda()
```

Out[55]:

	Results	Observations
Speedml Release	v0.9.3	Visit https://speedml.com for release notes.
Shape	train (19158, 14) | test (2129, 13)	
Numerical Ratio	28%	Aim for 100% numerical.
Numerical High-cardinality	[city_development_index, training_hours]	(>10) categories. Use feature.density
Numerical Categorical	[target]	Use plot.ordinal.
Numerical Continuous	[enrollee_id]	~80% unique. Use plot.continuous.
Text High-cardinality	[experience, city]	(>10) categories. Use feature.labels.
Text Categorical	[last_new_job, relevent_experience, gender, major_discipline, enrolled_university, education_level, company_size, company_type]	Use feature.labels or feature.mapping.
Target Analysis (target)	Model ready.	Use classification models.

In [56]:
```
sml.plot.strip('experience','training_hours')
```



In [57]:
```
sml.plot.bar('experience','training_hours')
```