trevor_eda

April 28, 2025

[1]:
```
from pandas import read_csv
kickstarter_filename = 'kickstarter_data_full.csv'
kickstarter_filename_features = 'kickstarter_data_with_features.csv'

ks_data = read_csv(kickstarter_filename)
ks_feat_data = read_csv(kickstarter_filename_features)
data_list = [('ks_data', ks_data), ('ks_feat_data', ks_feat_data)]
```

```
/tmp/ipykernel_6283/545315767.py:5: DtypeWarning: Columns (29,30,31,32) have
mixed types. Specify dtype option on import or set low_memory=False.
  ks_data = read_csv(kickstarter_filename)
/tmp/ipykernel_6283/545315767.py:6: DtypeWarning: Columns (29,30,31,32) have
mixed types. Specify dtype option on import or set low_memory=False.
  ks_feat_data = read_csv(kickstarter_filename_features)
```

[ ]:
```
from pandas.plotting import scatter_matrix
import matplotlib.pyplot as plt
import seaborn as sns
from pandas import set_option

def data_info(_data):
    print(f'Data head: {_data.head(5)}')
    print(f'Null values: {_data.isnull().sum()}')
    print(f'Data Shape: {_data.shape[0]} rows and {_data.shape[1]} columns')
    print(f'Columns: {list(_data.columns)}')
    for column in _data.columns:
        if len(_data[column].unique()) < 10:
            print(f'{column} unique values: {_data[column].unique()}')
        else:
            percent_unique = len(_data[column].unique()) / _data.shape[0] * 100
            print(f'{column} % unique values: {percent_unique}')
    print(_data.describe())
    _data.hist()
    plt.tight_layout()
    plt.show()
    plt.figure() # new plot
    plt.tight_layout()
    corMat = _data.corr(method='pearson')
```

```
    print(corMat)
    ## plot correlation matrix as a heat map
    sns.heatmap(corMat, square=True)
    plt.yticks(rotation=0)
    plt.xticks(rotation=90)
    plt.title(f"CORRELATION MATRIX USING HEAT MAP")
    plt.show()

    ## scatter plot of all _data
    plt.figure()
    # # The output overlaps itself, resize it to display better (w padding)
    scatter_matrix(_data)
    plt.tight_layout(pad=0.1)
    plt.show()

def data_info2(_date):
    set_option('display.max_columns', None)
    if not isinstance(_date, list):
        if not isinstance(_date, tuple):
            _date = ('', _date)
        _date = [_date]
    for name, data in _date:
        print(f'{name} data')
        print(data.info())
        print(data.head(5))
        for column in data.columns:
            highlight_column = 'profile'
            if column == highlight_column:
                # print 2 rows of values completely
                row1 = data.iloc[0][highlight_column]
                row2 = data.iloc[1][highlight_column]
                print(f'Row 1: {row1}')
                print(f'Row 2: {row2}')
            if len(data[column].unique()) < 10:
                print(f'{column} unique values: {data[column].unique()}')
            else:
                percent_unique = len(data[column].unique()) / data.shape[0] *␣
 ↪100
                print(f'{column} % unique values: {percent_unique}')
        break
#data_info2(ks_data)
columns_to_drop = ['Unnamed: 0', 'id', 'photo', 'name', 'blurb', 'slug',␣
 ↪'currency_symbol', 'currency_trailing_code', 'static_usd_rate', 'creator',␣
 ↪'profile', 'friends', 'is_backing', 'permissions', 'name_len', 'blurb_len',␣
 ↪'urls', 'source_url', 'location', 'is_starred', 'create_to_launch']
float_columns = ['goal', 'pledged', 'usd_pledged', ]
```

```
int_columns = ['backers_count', 'name_len_clean', 'blurb_len_clean', ␣
 ↪'launch_to_deadline', 'launch_to_state_change', 'create_to_launch_days',␣
 ↪'launch_to_deadline_days', 'launch_to_state_change_days', ]
datetime_columns = ['deadline', 'state_changed_at', 'created_at', 'launched_at']
date_int_columns = ['deadline_month', 'deadline_day', 'deadline_hr',␣
 ↪'state_changed_at_month', 'state_changed_at_day', 'state_changed_at_month',␣
 ↪'state_changed_at_day', 'state_changed_at_yr', 'created_at_month',␣
 ↪'created_at_day', 'created_at_hr', 'launched_at_month', 'launched_at_day',␣
 ↪'launched_at_hr', 'state_changed_at_hr', 'created_at_yr', 'launched_at_yr']
category_columns = ['state', 'currency', 'staff_pick', 'category',␣
 ↪'deadline_weekday', 'state_changed_at_weekday', 'created_at_weekday',␣
 ↪'launched_at_weekday', 'deadline_yr', 'country']
boolean_columns = ['disable_communication', 'spotlight', 'SuccessfulBool',␣
 ↪'USorGB', 'TOPCOUNTRY', 'LaunchedTuesday', 'DeadlineWeekend']

# This is just to categorize everything and make sure I'm not missing anything
temp_data = ks_data.copy()
temp_data = temp_data.drop(columns=columns_to_drop)
temp_data = temp_data.drop(columns=datetime_columns)
temp_data = temp_data.drop(columns=category_columns)
temp_data = temp_data.drop(columns=boolean_columns)
temp_data = temp_data.drop(columns=float_columns)
temp_data = temp_data.drop(columns=int_columns)
temp_data = temp_data.drop(columns=date_int_columns)
data_info2(('temp_data', temp_data))
kickstarter = ks_data.copy()
kickstarter = kickstarter.drop(columns=columns_to_drop)
```

[ ]:
```
# print columns not present in the other dataset
ks_data_columns = set(ks_data.columns)
ks_feat_data_columns = set(ks_feat_data.columns)

ks_data_not_in_feat = ks_data_columns - ks_feat_data_columns
ks_feat_data_not_in_ks = ks_feat_data_columns - ks_data_columns

print(f'Columns in ks_data not in ks_feat_data: {ks_data_not_in_feat}')
print(f'Columns in ks_feat_data not in ks_data: {ks_feat_data_not_in_ks}')

common_columns = ks_data_columns.intersection(ks_feat_data_columns)

# of common columns, compare the values and see if they match
for column in common_columns:
    ks_data_values = ks_data[column].unique()
    ks_feat_data_values = ks_feat_data[column].unique()
    if len(ks_data_values) != len(ks_feat_data_values):
        print(f'Column {column} has different number of unique values:␣
 ↪{len(ks_data_values)} vs {len(ks_feat_data_values)}')
```

```
    else:
        pass
```

[4]:
```
#using ks_data from here out
# convert name_len_clean and blurb_len_clean to int from float

# Create an explicit copy before making changes
ks_data = ks_data.copy()

# convert name_len_clean and blurb_len_clean to int from float
# show unique values to make sure we can convert them all to int
ks_data['name_len_clean'].unique()
# show me the row with 'nan' in it to guage what it should be converted to
ks_data[ks_data['name_len_clean'].isnull()]
# Nulls are for test data and 3 others, so just drop them
ks_data = ks_data.dropna(subset=['name_len_clean'])
ks_data['name_len_clean'] = ks_data['name_len_clean'].astype(int)
ks_data['blurb_len_clean'] = ks_data['blurb_len_clean'].astype(int)
```

[5]:
```
data_info2(('kickstarter', kickstarter))
```

```
kickstarter data
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20632 entries, 0 to 20631
Data columns (total 47 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   goal                         20632 non-null  float64
 1   pledged                      20632 non-null  float64
 2   state                        20632 non-null  object
 3   disable_communication        20632 non-null  bool
 4   country                      20632 non-null  object
 5   currency                     20632 non-null  object
 6   deadline                     20632 non-null  object
 7   state_changed_at             20632 non-null  object
 8   created_at                   20632 non-null  object
 9   launched_at                  20632 non-null  object
 10  staff_pick                   20632 non-null  bool
 11  backers_count                20632 non-null  int64
 12  usd_pledged                  20632 non-null  float64
 13  category                     18743 non-null  object
 14  spotlight                    20632 non-null  bool
 15  name_len_clean               20627 non-null  float64
 16  blurb_len_clean              20627 non-null  float64
 17  deadline_weekday             20632 non-null  object
 18  state_changed_at_weekday     20632 non-null  object
 19  created_at_weekday           20632 non-null  object
 20  launched_at_weekday          20632 non-null  object
```

```
 21  deadline_month               20632 non-null  int64
 22  deadline_day                 20632 non-null  int64
 23  deadline_yr                  20632 non-null  int64
 24  deadline_hr                  20632 non-null  int64
 25  state_changed_at_month       20632 non-null  int64
 26  state_changed_at_day         20632 non-null  int64
 27  state_changed_at_yr          20632 non-null  int64
 28  state_changed_at_hr          20632 non-null  int64
 29  created_at_month             20632 non-null  int64
 30  created_at_day               20632 non-null  int64
 31  created_at_yr                20632 non-null  int64
 32  created_at_hr                20632 non-null  int64
 33  launched_at_month            20632 non-null  int64
 34  launched_at_day              20632 non-null  int64
 35  launched_at_yr               20632 non-null  int64
 36  launched_at_hr               20632 non-null  int64
 37  launch_to_deadline           20632 non-null  object
 38  launch_to_state_change       20632 non-null  object
 39  create_to_launch_days        20632 non-null  int64
 40  launch_to_deadline_days      20632 non-null  int64
 41  launch_to_state_change_days  20632 non-null  int64
 42  SuccessfulBool               20632 non-null  int64
 43  USorGB                       20632 non-null  int64
 44  TOPCOUNTRY                   20632 non-null  int64
 45  LaunchedTuesday              20632 non-null  int64
 46  DeadlineWeekend              20632 non-null  int64
dtypes: bool(3), float64(5), int64(25), object(14)
memory usage: 7.0+ MB
None
       goal  pledged   state  disable_communication country currency  \
0    1500.0      0.0  failed                  False      US      USD
1     500.0      0.0  failed                  False      US      USD
2  100000.0    120.0  failed                  False      US      USD
3    5000.0      0.0  failed                  False      US      USD
4    3222.0    356.0  failed                  False      DE      EUR

          deadline state_changed_at        created_at       launched_at  \
0  1/23/2015 10:35  1/23/2015 10:35  11/29/2014 22:55  12/17/2014 13:47
1   5/1/2015 16:13   5/1/2015 16:13    2/20/2015 9:28    3/2/2015 16:13
2   3/26/2015 8:17   3/26/2015 8:17    1/24/2015 0:08    1/25/2015 8:17
3   10/6/2014 0:41   10/6/2014 0:41    9/5/2014 22:30     9/6/2014 0:41
4  6/27/2016 12:00  6/27/2016 12:00   5/25/2016 14:09    5/26/2016 5:57

   staff_pick  backers_count  usd_pledged  category  spotlight  \
0       False              0     0.000000  Academic      False
1       False              0     0.000000  Academic      False
2       False              5   120.000000  Academic      False
3       False              0     0.000000  Academic      False
```

```
4       False             17   396.802395  Academic      False

   name_len_clean  blurb_len_clean deadline_weekday state_changed_at_weekday  \
0             9.0             16.0           Friday                   Friday
1             4.0             15.0           Friday                   Friday
2             8.0             10.0         Thursday                 Thursday
3             6.0             13.0           Monday                   Monday
4             7.0             18.0           Monday                   Monday

  created_at_weekday launched_at_weekday  deadline_month  deadline_day  \
0           Saturday           Wednesday               1            23
1             Friday              Monday               5             1
2           Saturday              Sunday               3            26
3             Friday            Saturday              10             6
4          Wednesday            Thursday               6            27

   deadline_yr  deadline_hr  state_changed_at_month  state_changed_at_day  \
0         2015           10                       1                    23
1         2015           16                       5                     1
2         2015            8                       3                    26
3         2014            0                      10                     6
4         2016           12                       6                    27

   state_changed_at_yr  state_changed_at_hr  created_at_month  created_at_day  \
0                 2015                   10                11              29
1                 2015                   16                 2              20
2                 2015                    8                 1              24
3                 2014                    0                 9               5
4                 2016                   12                 5              25

   created_at_yr  created_at_hr  launched_at_month  launched_at_day  \
0           2014             22                 12               17
1           2015              9                  3                2
2           2015              0                  1               25
3           2014             22                  9                6
4           2016             14                  5               26

   launched_at_yr  launched_at_hr          launch_to_deadline  \
0            2014              13  36 days 20:47:24.000000000
1            2015              16  60 days 00:00:00.000000000
2            2015               8  60 days 00:00:00.000000000
3            2014               0  30 days 00:00:00.000000000
4            2016               5  32 days 06:02:33.000000000

       launch_to_state_change  create_to_launch_days  launch_to_deadline_days  \
0  36 days 20:47:24.000000000                     17                       36
1  60 days 00:00:02.000000000                     10                       60
2  60 days 00:00:01.000000000                      1                       60
```

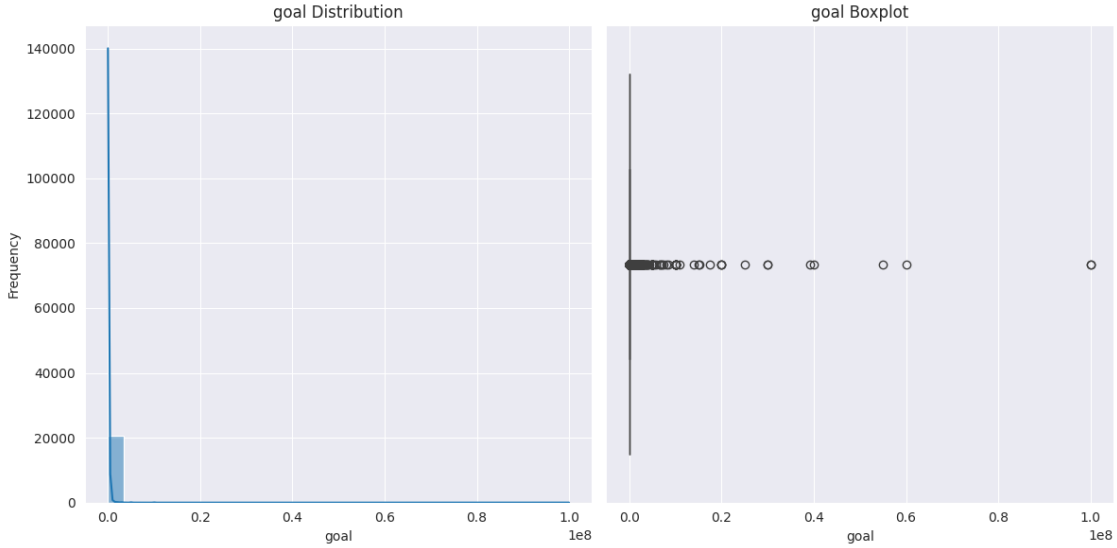
```
3  30 days 00:00:00.000000000                     0                      30
4  32 days 06:02:33.000000000                     0                      32

   launch_to_state_change_days  SuccessfulBool  USorGB  TOPCOUNTRY  \
0                           36               0       1           1
1                           60               0       1           1
2                           60               0       1           1
3                           30               0       1           1
4                           32               0       0           0

   LaunchedTuesday  DeadlineWeekend
0                0                0
1                0                0
2                0                0
3                0                0
4                0                0
goal % unique values: 5.835595191934859
pledged % unique values: 42.274137262504844
state unique values: ['failed' 'canceled' 'successful' 'live' 'suspended']
disable_communication unique values: [False  True]
country % unique values: 0.10178363706863124
currency % unique values: 0.06300891818534315
deadline % unique values: 97.96917409848778
state_changed_at % unique values: 98.09034509499807
created_at % unique values: 99.27782086079876
launched_at % unique values: 99.10818146568438
staff_pick unique values: [False  True]
backers_count % unique values: 7.294493989918573
usd_pledged % unique values: 59.30108569212873
category % unique values: 0.1211709965102753
spotlight unique values: [False  True]
name_len_clean % unique values: 0.07270259790616518
blurb_len_clean % unique values: 0.14540519581233036
deadline_weekday unique values: ['Friday' 'Thursday' 'Monday' 'Sunday' 'Tuesday'
'Wednesday' 'Saturday']
state_changed_at_weekday unique values: ['Friday' 'Thursday' 'Monday' 'Sunday'
'Tuesday' 'Wednesday' 'Saturday']
created_at_weekday unique values: ['Saturday' 'Friday' 'Wednesday' 'Monday'
'Thursday' 'Sunday' 'Tuesday']
launched_at_weekday unique values: ['Wednesday' 'Monday' 'Sunday' 'Saturday'
'Thursday' 'Tuesday' 'Friday']
deadline_month % unique values: 0.05816207832493214
deadline_day % unique values: 0.15025203567274137
deadline_yr unique values: [2015 2014 2016 2017 2011 2013 2012 2010 2009]
deadline_hr % unique values: 0.11632415664986429
state_changed_at_month % unique values: 0.05816207832493214
state_changed_at_day % unique values: 0.15025203567274137
state_changed_at_yr unique values: [2015 2014 2016 2017 2011 2013 2012 2010
```

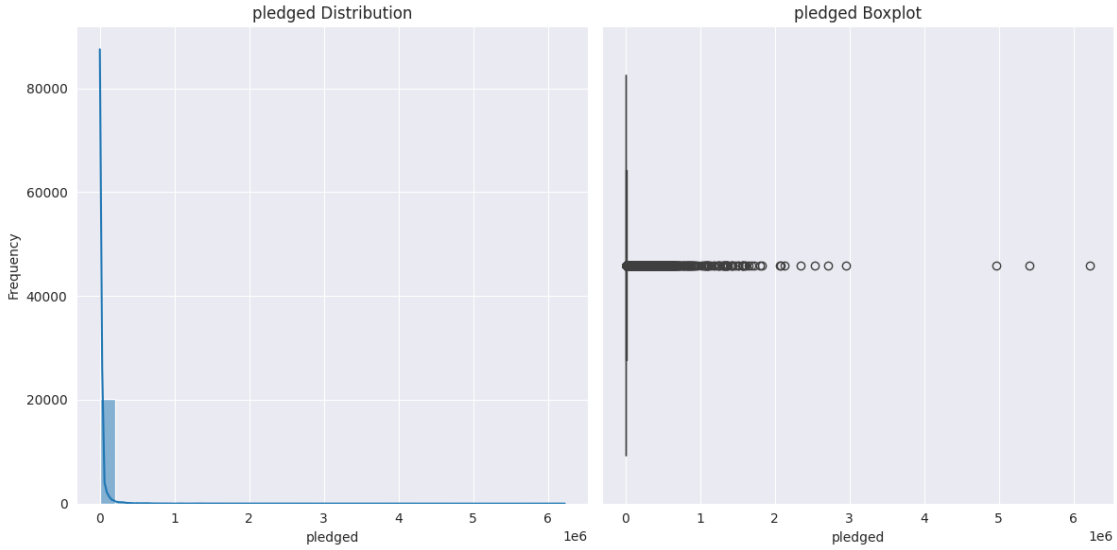
```
2009]
state_changed_at_hr % unique values: 0.11632415664986429
created_at_month % unique values: 0.05816207832493214
created_at_day % unique values: 0.15025203567274137
created_at_yr unique values: [2014 2015 2016 2017 2013 2012 2011 2010 2009]
created_at_hr % unique values: 0.11632415664986429
launched_at_month % unique values: 0.05816207832493214
launched_at_day % unique values: 0.15025203567274137
launched_at_yr unique values: [2014 2015 2016 2017 2011 2013 2012 2010 2009]
launched_at_hr % unique values: 0.11632415664986429
launch_to_deadline % unique values: 24.171190383869718
launch_to_state_change % unique values: 36.56455990694067
create_to_launch_days % unique values: 3.3249321442419544
launch_to_deadline_days % unique values: 0.41198138813493607
launch_to_state_change_days % unique values: 0.40713454827452494
SuccessfulBool unique values: [0 1]
USorGB unique values: [1 0]
TOPCOUNTRY unique values: [1 0]
LaunchedTuesday unique values: [0 1]
DeadlineWeekend unique values: [0 1]
```

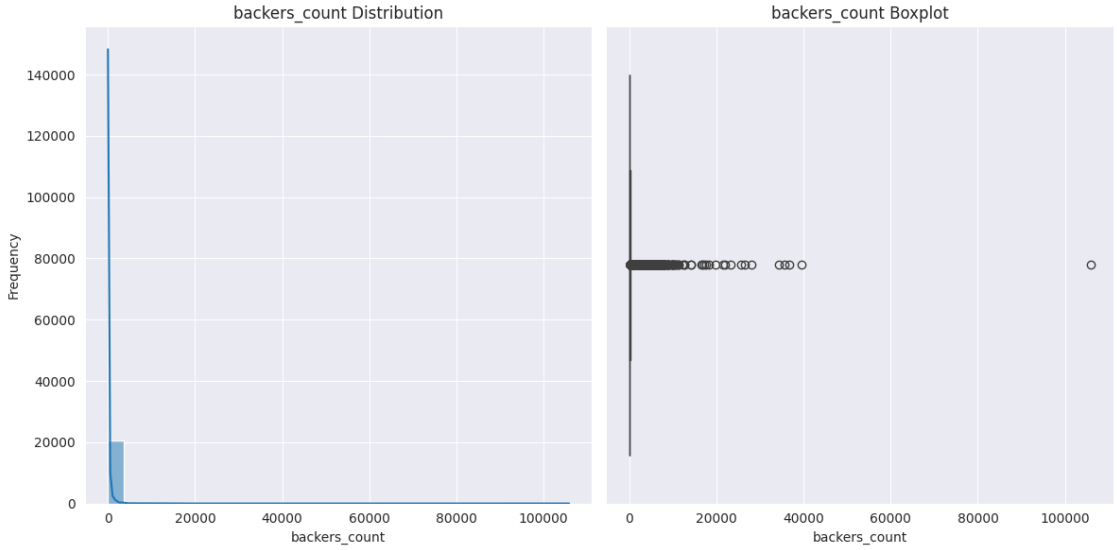
0.1 Now we know what is in the data, let's look at some of the features and see if we can find any interesting patterns.

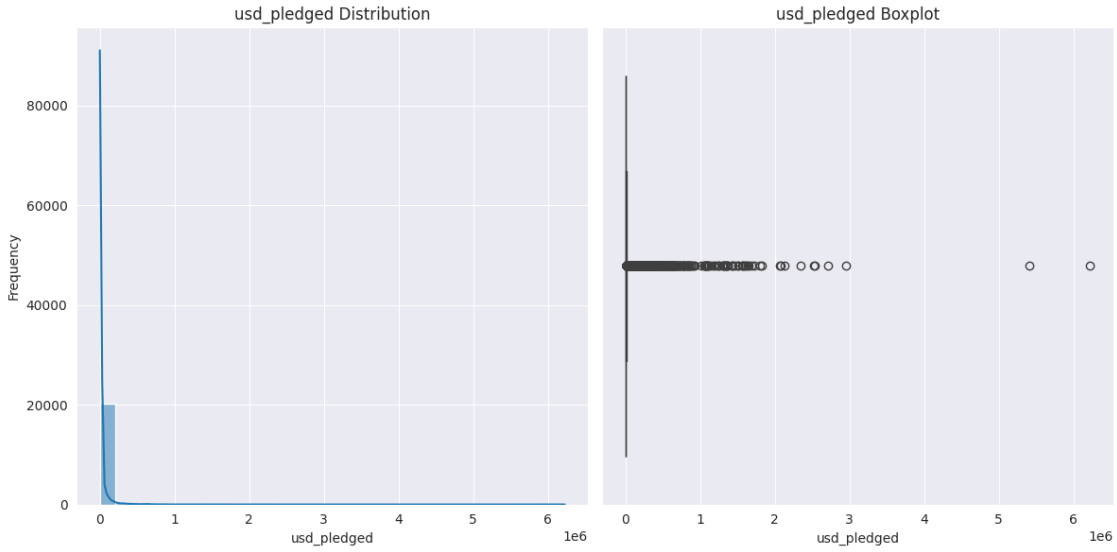
[6]:
```
# For each feature, lets see a histogram of the values and a boxplot of the values

def plot_feature_distribution(data, feature):
    plt.figure(figsize=(12, 6))
    plt.subplot(1, 2, 1)
    sns.histplot(data[feature], bins=30, kde=True)
    plt.title(f'{feature} Distribution')
    plt.xlabel(feature)
    plt.ylabel('Frequency')

    plt.subplot(1, 2, 2)
    sns.boxplot(x=data[feature])
    plt.title(f'{feature} Boxplot')
    plt.xlabel(feature)

    plt.tight_layout()
    plt.show()

columns_to_show = float_columns + int_columns + date_int_columns

for feature in kickstarter.columns:
    if feature in columns_to_show:
```

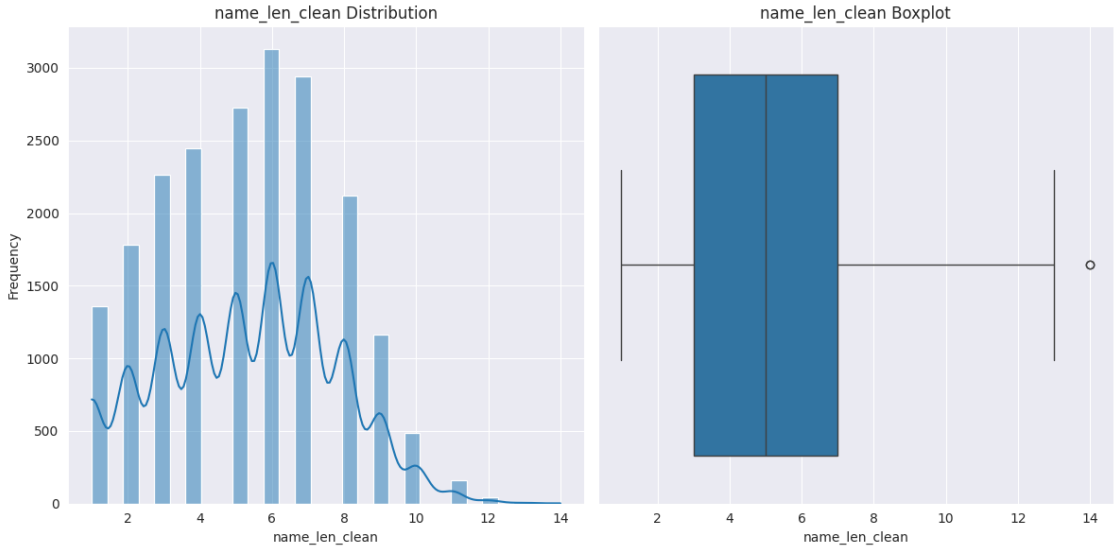
```
plot_feature_distribution(kickstarter, feature)
```

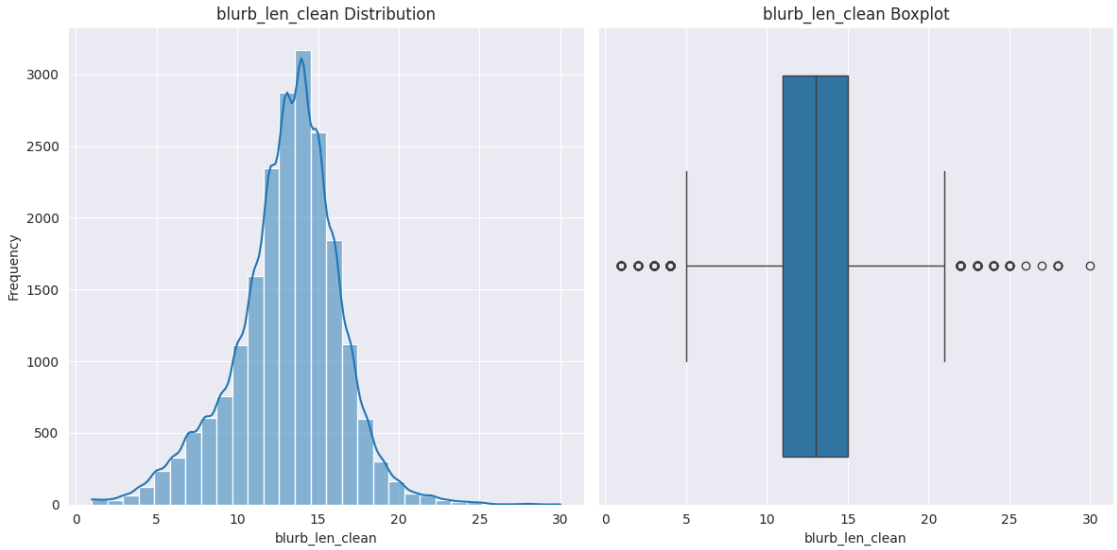


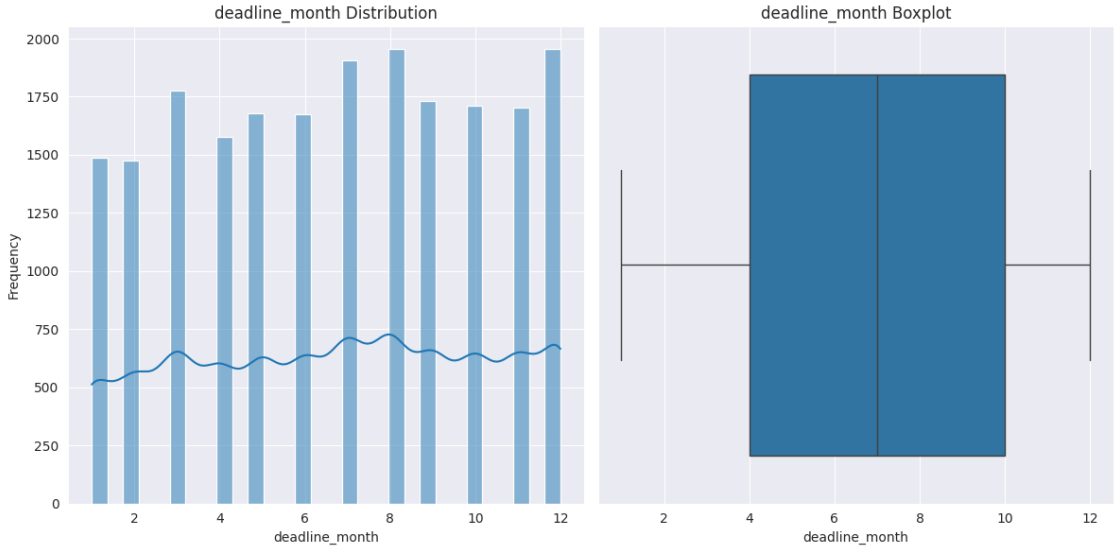


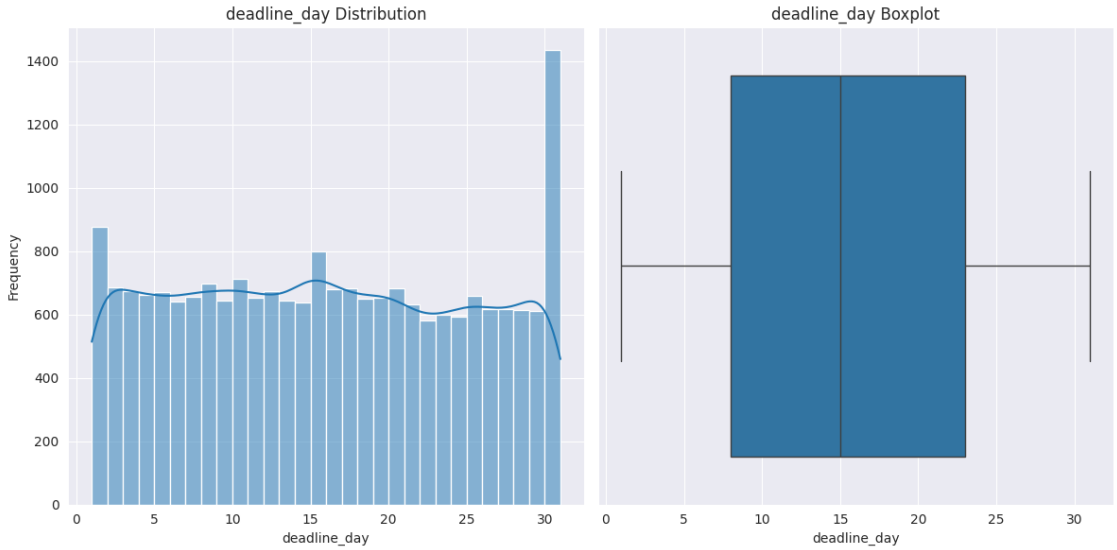

backers_count Distribution
backers_count Boxplot
Frequency
backers_count
backers_count

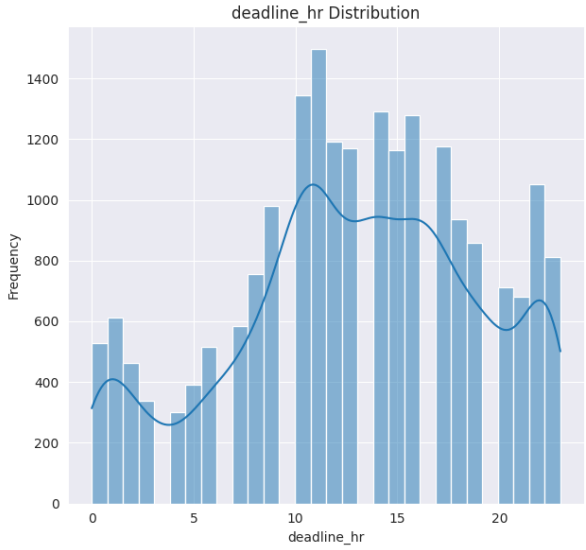

usd_pledged Distribution
usd_pledged Boxplot
Frequency
usd_pledged
1e6
usd_pledged
1e6

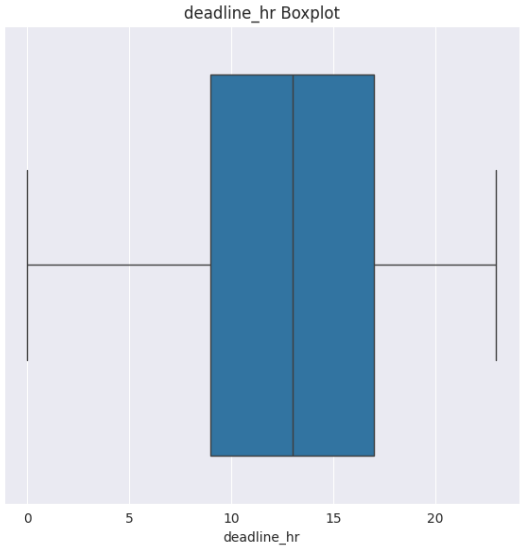

name_len_clean Distribution
Frequency
name_len_clean
name_len_clean Boxplot
name_len_clean

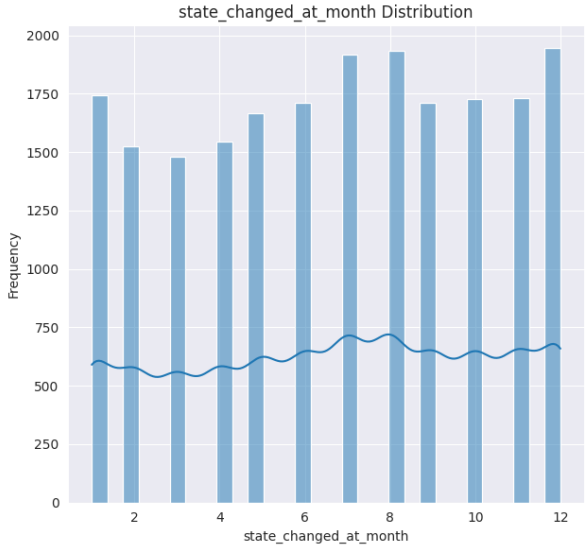

blurb_len_clean Distribution
Frequency
blurb_len_clean
blurb_len_clean Boxplot
blurb_len_clean

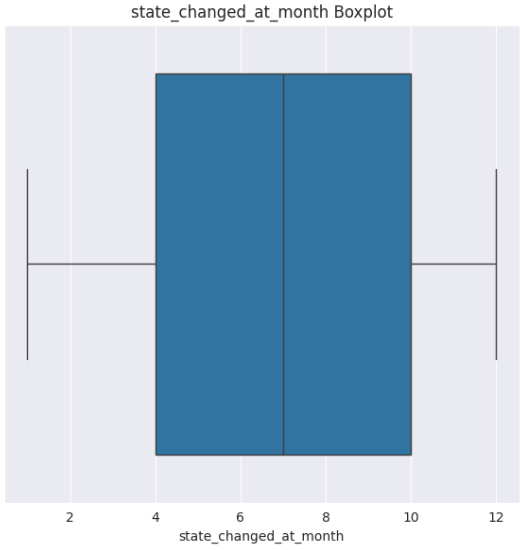

deadline_month Distribution
Frequency
deadline_month
deadline_month Boxplot
deadline_month

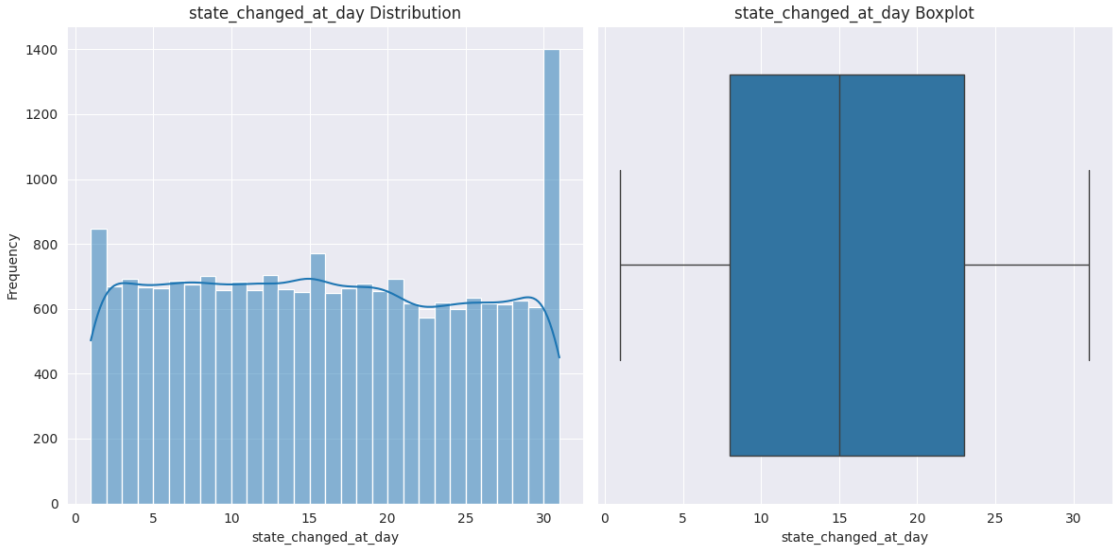

deadline_day Distribution
Frequency
deadline_day
deadline_day Boxplot
deadline_day

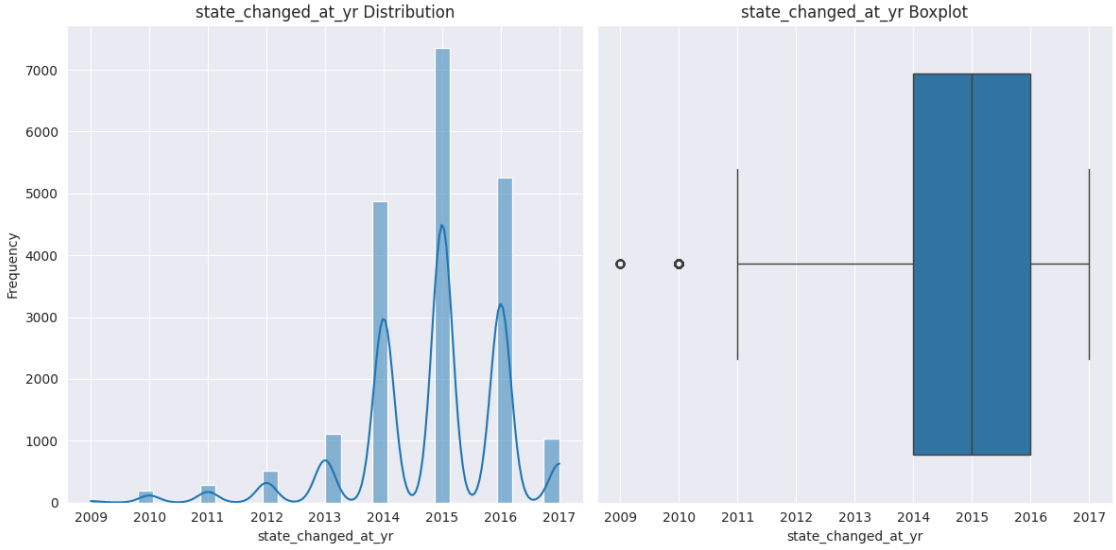

deadline_hr Distribution
Frequency
deadline_hr
0
200
400
600
800
1000
1200
1400
0
5
10
15
20

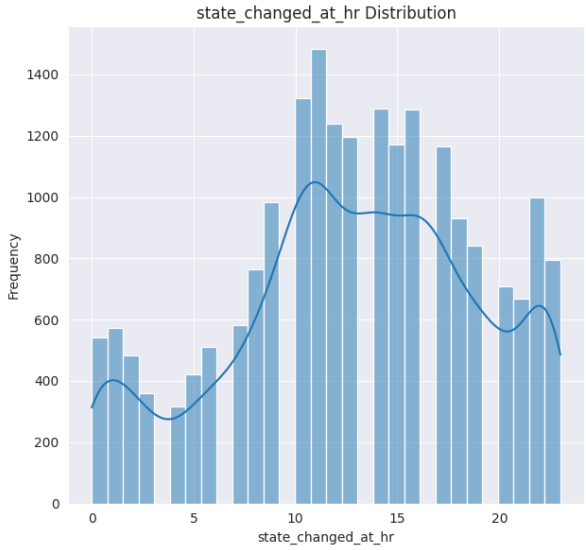

deadline_hr Boxplot
deadline_hr
0
5
10
15
20

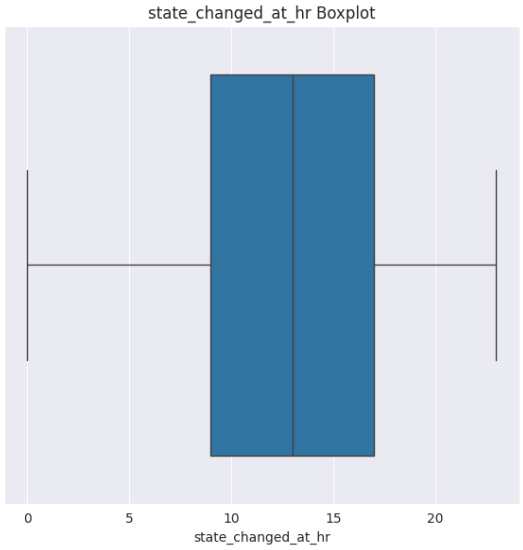

state_changed_at_month Distribution
Frequency
state_changed_at_month
0
250
500
750
1000
1250
1500
1750
2000
2
4
6
8
10
12

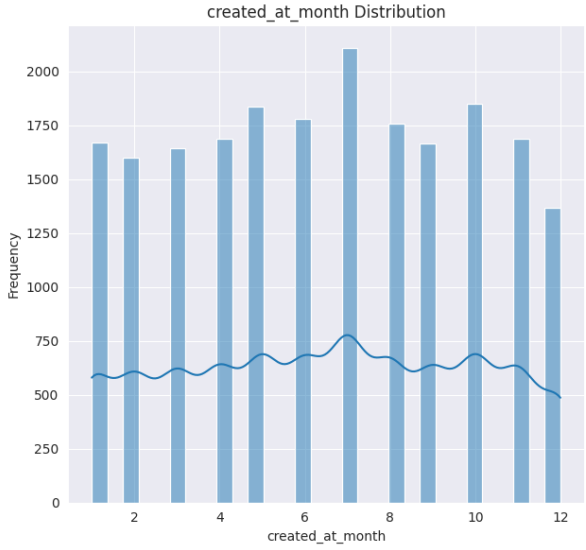

state_changed_at_month Boxplot
state_changed_at_month
2
4
6
8
10
12

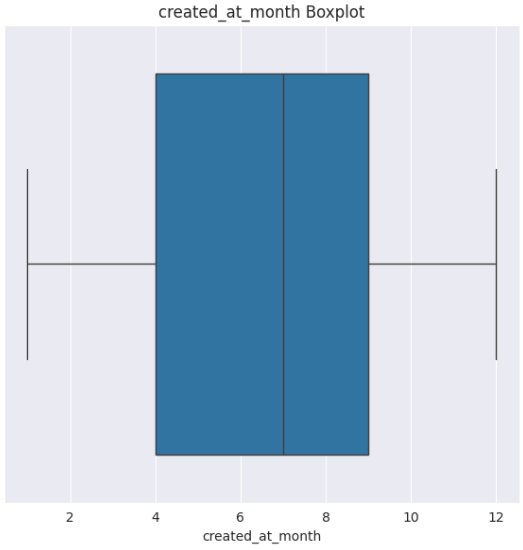

state_changed_at_day Distribution
Frequency
state_changed_at_day
state_changed_at_day Boxplot
state_changed_at_day

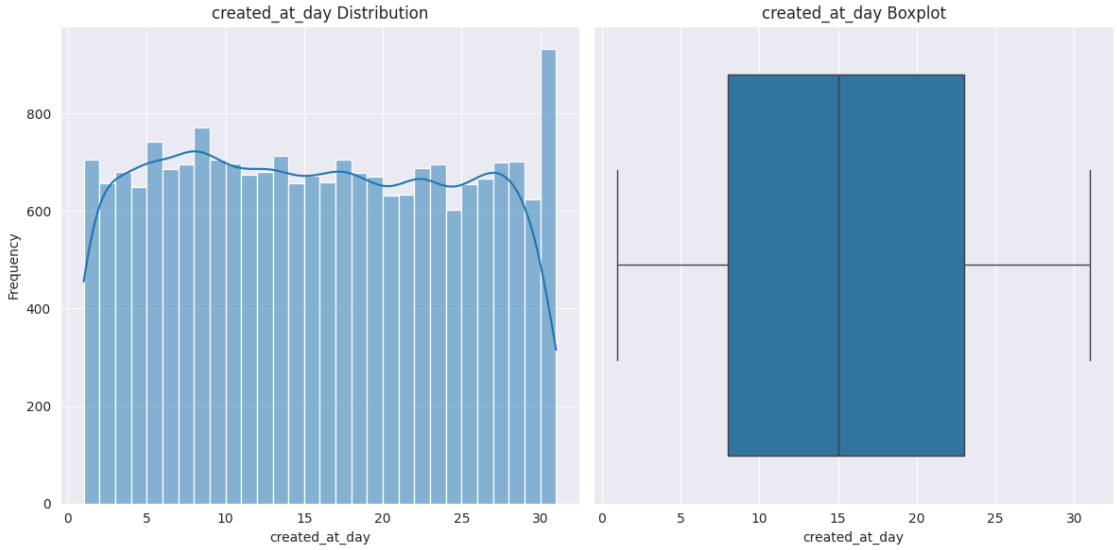

state_changed_at_yr Distribution
Frequency
state_changed_at_yr
state_changed_at_yr Boxplot
state_changed_at_yr

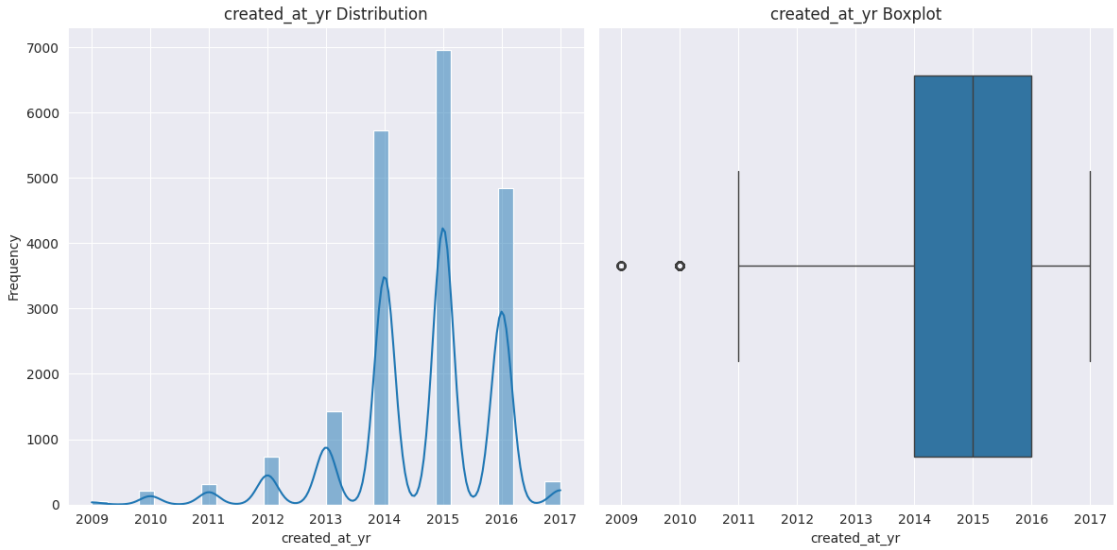

state_changed_at_hr Distribution
Frequency
state_changed_at_hr
0
200
400
600
800
1000
1200
1400
0
5
10
15
20

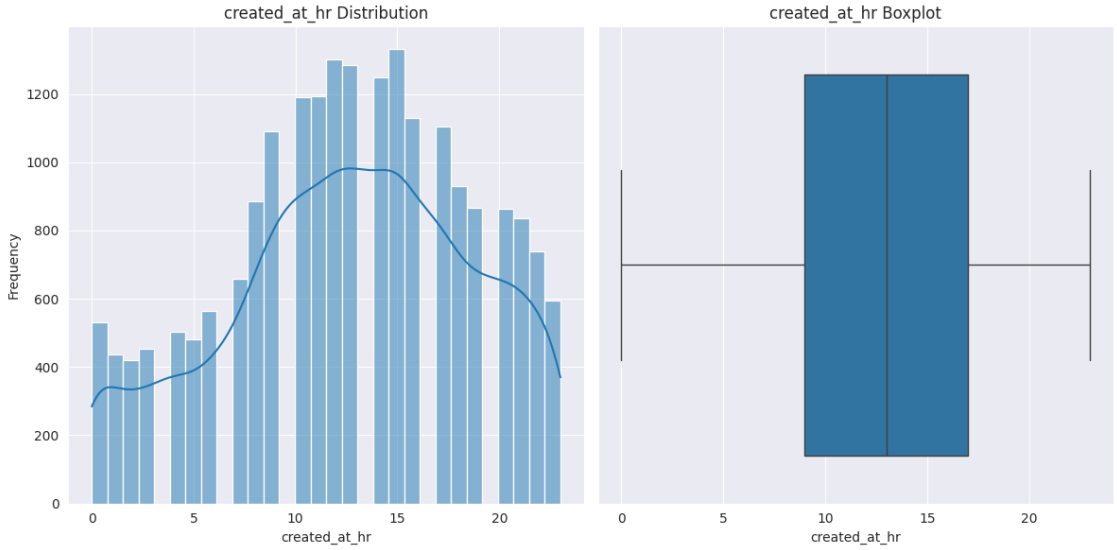

state_changed_at_hr Boxplot
state_changed_at_hr
0
5
10
15
20

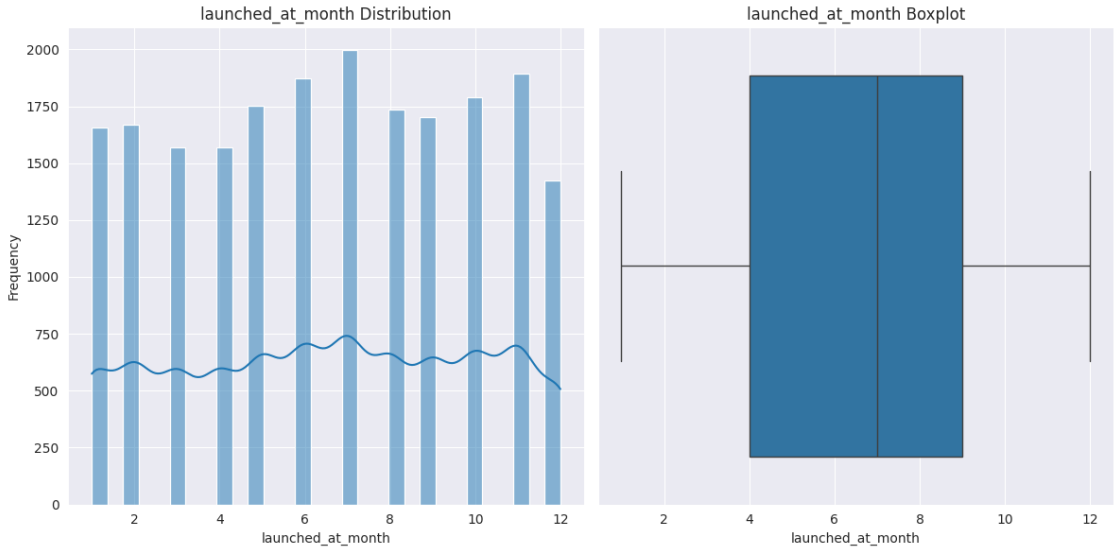

created_at_month Distribution
Frequency
created_at_month
0
250
500
750
1000
1250
1500
1750
2000
2
4
6
8
10
12

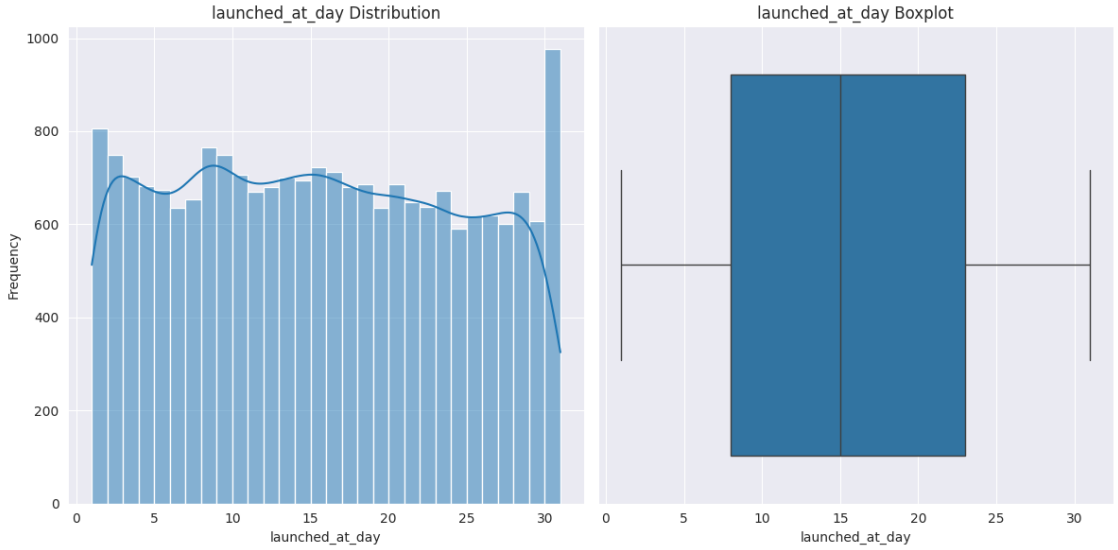

created_at_month Boxplot
created_at_month
2
4
6
8
10
12

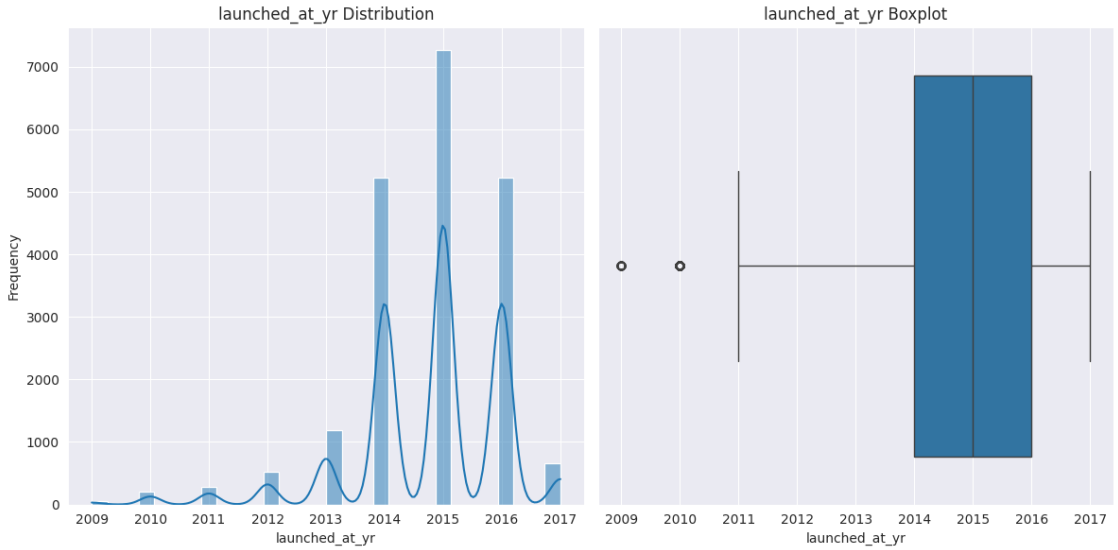

created_at_day Distribution
Frequency
created_at_day
created_at_day Boxplot
created_at_day

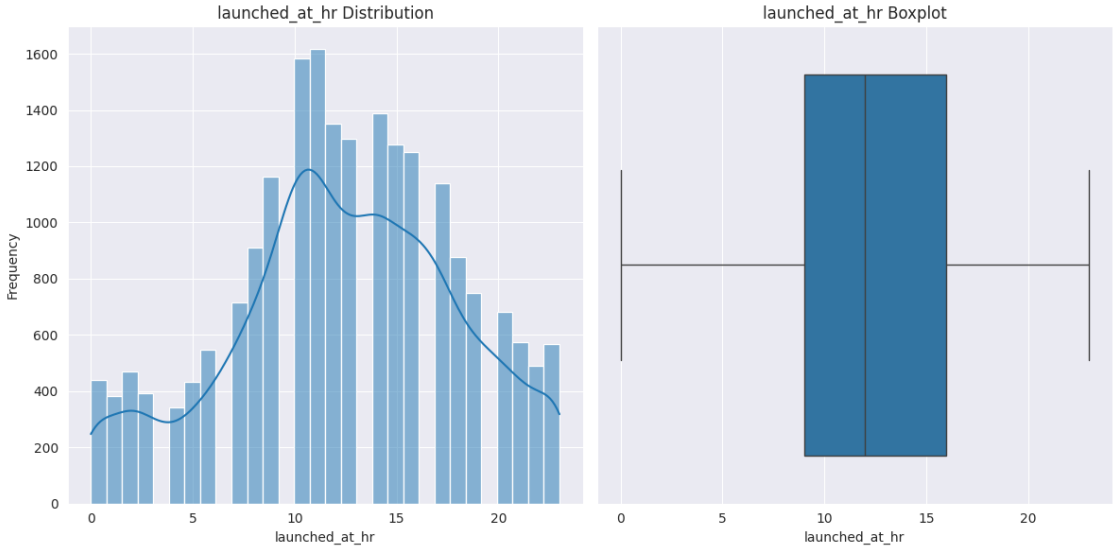

created_at_yr Distribution
Frequency
created_at_yr
created_at_yr Boxplot
created_at_yr

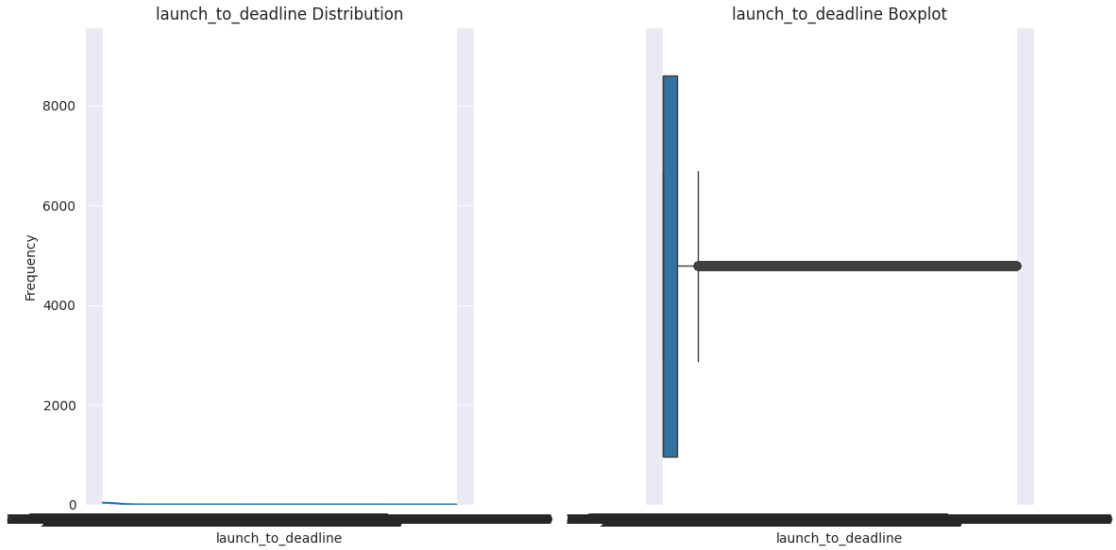

created_at_hr Distribution
Frequency
created_at_hr
created_at_hr Boxplot
created_at_hr

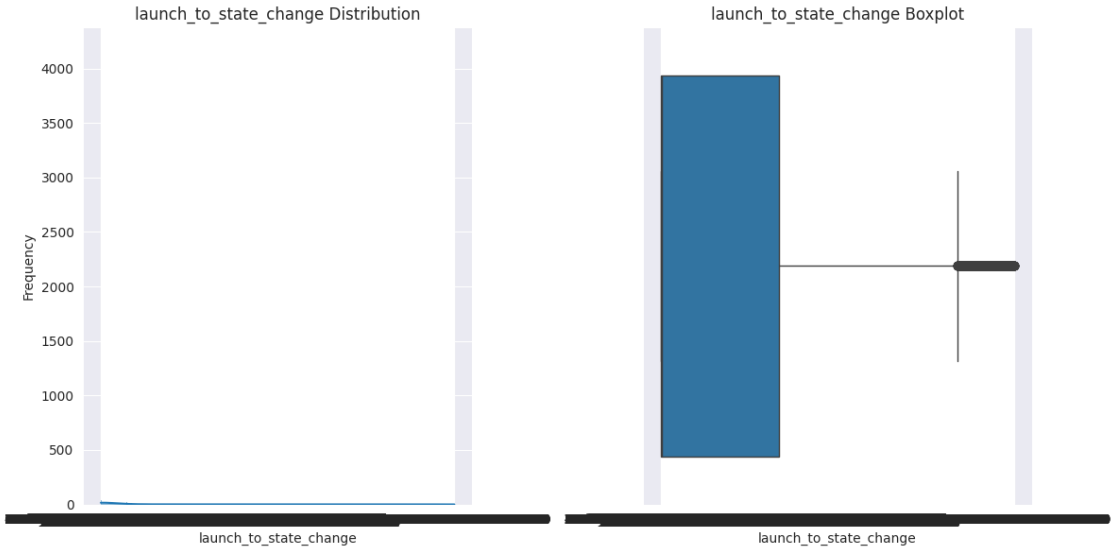

launched_at_month Distribution
Frequency
launched_at_month
launched_at_month Boxplot
launched_at_month

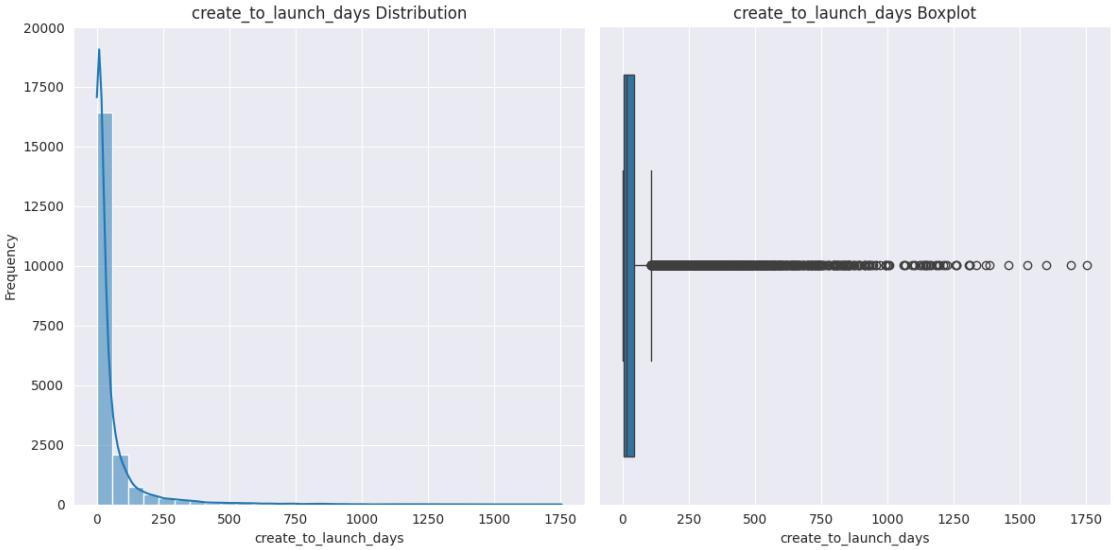

launched_at_day Distribution
Frequency
launched_at_day
launched_at_day Boxplot
launched_at_day

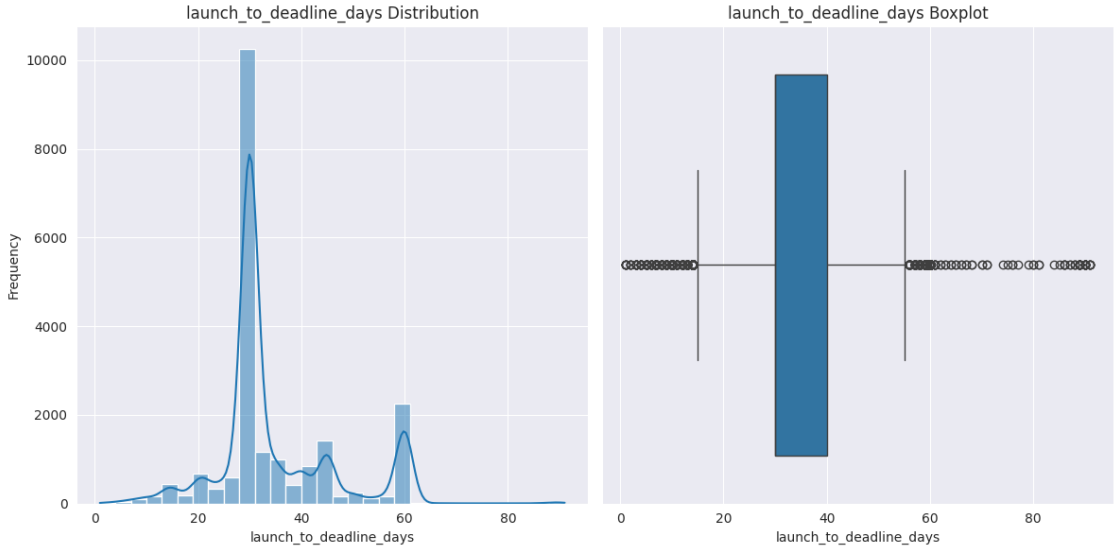

launched_at_yr Distribution
Frequency
launched_at_yr
launched_at_yr Boxplot
launched_at_yr

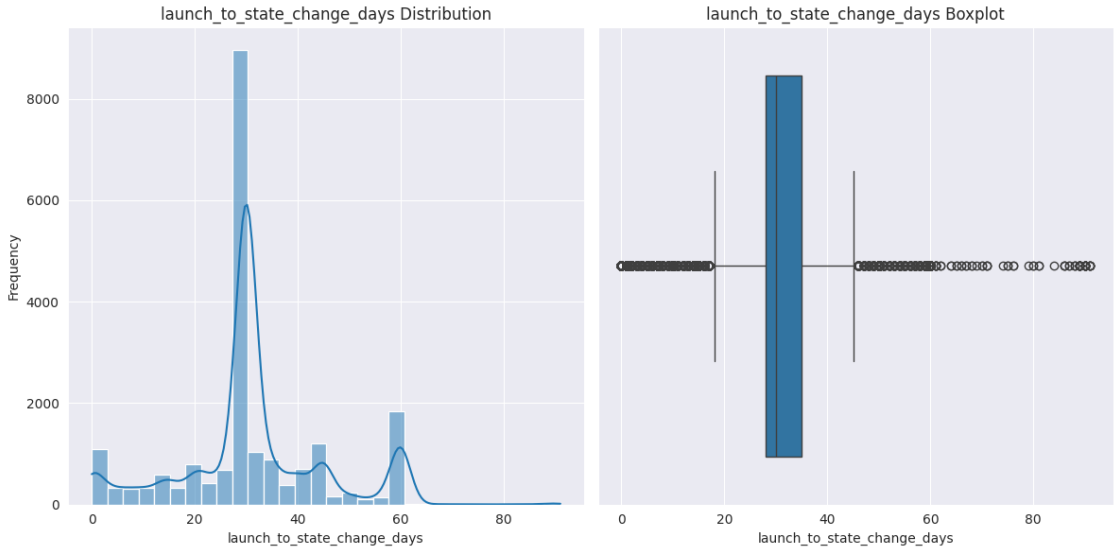

launched_at_hr Distribution
Frequency
launched_at_hr
launched_at_hr Boxplot
launched_at_hr


launch_to_deadline Distribution
Frequency
launch_to_deadline
launch_to_deadline Boxplot
launch_to_deadline


launch_to_state_change Distribution
launch_to_state_change Boxplot
Frequency
launch_to_state_change
launch_to_state_change


create_to_launch_days Distribution
create_to_launch_days Boxplot
Frequency
create_to_launch_days
create_to_launch_days


launch_to_deadline_days Distribution
Frequency
launch_to_deadline_days
launch_to_deadline_days Boxplot
launch_to_deadline_days


launch_to_state_change_days Distribution
Frequency
launch_to_state_change_days
launch_to_state_change_days Boxplot
launch_to_state_change_days