Proyecto

Carrela Silva Carlos Francisco Javier

Importar librerias

In [1]:
```
import pandas as pd
import numpy as np
import pandas as pd
import numpy as np
# Statistics
from scipy.stats import ks_2samp
from scipy.stats import chisquare
#visualization
import cufflinks as cf
#reduccion de variables
from sklearn.pipeline import Pipeline
from sklearn.impute import SimpleImputer
from sklearn.preprocessing import MinMaxScaler
from sklearn.model_selection import train_test_split, cross_val_score
from sklearn.decomposition import PCA
from sklearn.feature_selection import SelectKBest, f_regression
from sklearn.feature_selection import SelectKBest, f_classif
from sklearn.model_selection import train_test_split # pip install scikit-learn
from sklearn.preprocessing import KBinsDiscretizer
from sklearn.preprocessing import MinMaxScaler, RobustScaler, StandardScaler
from varclushi import VarClusHi
import matplotlib.pyplot as plt
```

La intencion de nuestro trabajo sera crear un modelo que nos ayude a calcular la apuesta Over de 2 goles de un determinado partido

Lectura de datos

In [2]:
```
club_games=pd.read_csv("/home/carlos/Documentos/diplomado/modulo 1/Proyecto/club_games.csv")
clubs=pd.read_csv("/home/carlos/Documentos/diplomado/modulo 1/Proyecto/clubs.csv")
game_events=pd.read_csv("/home/carlos/Documentos/diplomado/modulo 1/Proyecto/game_events.csv")
games=pd.read_csv("/home/carlos/Documentos/diplomado/modulo 1/Proyecto/games.csv")
```

In [3]:
```
club_games.sample(5)
```

Out[3]:

	club_id	game_id	own_goals	own_position	own_manager_name	opponent_id	opponent_goals	opponent_position	opponent_
29410	415	2942864	0	17	Mickaël Debève	2969	1	11	
113349	583	3589569	3	1	Mauricio Pochettino	667	3	6	
29503	1420	2942799	0	18	Stéphane Moulin	162	4	2	
103417	2578	3201328	0	12	Tommy Wright	987	4	4	St
41187	28643	3237907	0	16	Arnauld Mercier	172	4	10	

In [4]:
```
clubs.sample(5)
```

Out[4]:

	club_id	club_code	name	domestic_competition_id	total_market_value	squad_size	average_age	foreigners_number	foreigr
265	26459	nk-veres-rivne	Nk Veres Rivne	UKR1	NaN	23	26.8	0	
352	1420	sco-angers	Sco Angers	FR1	NaN	31	25.0	15	
205	2919	ac-monza	Ac Monza	IT1	NaN	30	26.6	9	
71	2944	ankaraspor	Ankaraspor	TR1	2.75	30	25.5	2	
131	2969	dijon-fco	Dijon Fco	FR1	16.61	24	27.0	11	

In [5]:
```
game_events.sample(5)
```

Out[5]:

	game_id	minute	type	club_id	player_id	description	player_in_id
273655	2894321	72	Substitutions	8838	171730	NaN	540186
45236	2235792	66	Goals	4772	75665	, 1. Tournament Goal	-1
39298	2251572	67	Substitutions	130	45567	NaN	46062
433579	3535921	93	Substitutions	2671	195704	NaN	494679
44542	2287211	74	Substitutions	46	25441	NaN	164534

In [6]:
```
games.sample(5)
```

Out[6]:

	game_id	competition_id	competition_type	season	round	date	home_club_id	away_club_id	home_club_goals	away_club
10354	2359370	IT1	domestic_league	2013	17. Matchday	2013-12-21	1210	410	1	
768	2229838	FR1	domestic_league	2012	2. Matchday	2012-08-18	1041	1095	4	
44070	3244632	ELQ	international_cup	2019	Qualifying Round 1st leg	2019-08-22	31614	790	0	
18226	2609421	PO1	domestic_league	2015	12. Matchday	2015-12-05	1301	336	0	
37773	3110066	RU1	domestic_league	2018	10. Matchday	2018-10-06	1083	14589	0	

5 rows × 21 columns

Agrupar y unir las tablas

In [7]:
```
grouped_game_goals= game_events.groupby(["game_id"]).agg({'type': lambda x: sum(x == 'Goals')}).sort_values('game_i
grouped_game_substitutions= game_events.groupby(["game_id"]).agg({'type': lambda x: sum(x == 'Substitutions')}).sor
goals_addedtime=game_events.groupby(["game_id"]).agg({'minute': lambda x: sum(x >90)}).sort_values('game_id') #esta
minute_last_goal = game_events.groupby('game_id').apply(lambda x: x.loc[x['type'] == 'Goals', 'minute'].max()).fill
minute_first_goal = game_events.groupby('game_id').apply(lambda x: x.loc[x['type'] == 'Goals', 'minute'].min()).fil
```

In [8]:
```
events=grouped_game_goals
events["number_game_substitutions"]=grouped_game_substitutions["type"]
events["goals_added_time"]=goals_addedtime["minute"]
events["minute_last_goal"]=minute_last_goal.set_index("game_id")[0]
events["minute_first_goal"]=minute_first_goal.set_index("game_id")[0]
events = events.rename(columns={'type': 'number_goals'})
events
```

Out[8]:

game_id	number_goals	number_game_substitutions	goals_added_time	minute_last_goal	minute_first_goal
2211607	6	5	1	94.0	3.0
2218677	6	6	4	101.0	27.0
2219794	5	6	0	76.0	24.0
2219795	5	6	0	74.0	6.0
2221641	7	6	0	83.0	27.0
...	...	...	...	...	...
4034848	2	8	0	67.0	21.0
4035007	5	10	0	90.0	6.0
4035008	6	10	0	73.0	32.0
4035009	4	6	1	96.0	50.0
4035010	5	7	4	91.0	21.0

61060 rows × 5 columns

In [9]:
```
events.describe()
```

Out[9]:

	number_goals	number_game_substitutions	goals_added_time	minute_last_goal	minute_first_goal
count	61060.000000	61060.000000	61060.000000	61060.000000	61060.000000
mean	2.766590	6.337848	0.416345	65.952080	29.311988
std	1.698559	1.587827	0.835234	27.818171	24.126591
min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	2.000000	6.000000	0.000000	53.000000	10.000000
50%	3.000000	6.000000	0.000000	75.000000	23.000000
75%	4.000000	6.000000	1.000000	87.000000	45.000000
max	17.000000	15.000000	11.000000	125.000000	122.000000

In [10]:
```
club_games # de esta tabla podemos conservar todo si solo tomamos como referencia a los partidos de locales, de toa
```

Out[10]:

	club_id	game_id	own_goals	own_position	own_manager_name	opponent_id	opponent_goals	opponent_position	opponent_
0	27	2229332	2	-1	Jupp Heynckes	16	1	-1	
1	131	2244388	3	-1	Tito Vilanova	418	2	-1	
2	3709	2269557	0	-1	Luis García	4032	0	-1	C
3	21322	2254432	1	-1	Pedro Buenaventura	7077	0	-1	
4	109	2221759	0	-1	Oscar Corrochano	27	4	-1	
...	...	...	...	...	...	...	...	...	
122343	865	3828462	1	8	Albert Capellas	2778	1	1	Fle
122344	865	3828427	5	5	Henrik Jensen	173	1	12	
122345	1063	3828419	1	6	Jacob Friis	678	3	5	
122346	1177	3828414	2	6	Kent Nielsen	369	2	5	Frey
122347	2414	3828447	0	7	Jens Berthel Askou	173	1	11	

122348 rows × 11 columns

In [11]:
```
#la tabla games esta lista para usar
games
```

Out[11]:

	game_id	competition_id	competition_type	season	round	date	home_club_id	away_club_id	home_club_goals	away_club
0	2229332	DFL	other	2012	Final	2012-08-12	27	16	2	
1	2244388	SUC	other	2012	final 1st leg	2012-08-22	131	418	3	
2	2269557	CDR	domestic_cup	2012	4th round 2nd leg	2012-11-28	3709	4032	0	
3	2254432	CDR	domestic_cup	2012	First Round Replay	2012-08-30	21322	7077	1	
4	2221759	DFB	domestic_cup	2012	First Round	2012-08-20	109	27	0	
...	...	...	...	...	...	...	...	...	...	
61169	3828462	DK1	domestic_league	2022	9. Matchday	2022-09-11	2778	865	1	
61170	3828427	DK1	domestic_league	2021	3. Matchday	2022-07-29	173	865	1	
61171	3828419	DK1	domestic_league	2021	2. Matchday	2022-07-24	678	1063	3	
61172	3828414	DK1	domestic_league	2021	1. Matchday	2022-07-17	369	1177	2	
61173	3828447	DK1	domestic_league	2022	6. Matchday	2022-08-22	173	2414	1	

61174 rows × 21 columns

In [12]:
```
merge_table= pd.merge(games.set_index("game_id"), club_games.set_index("game_id"), left_index=True, right_index=Tru
merge_table=pd.merge(merge_table, events, left_index=True, right_index=True,how="inner")
```

In [13]:
```
# para agregar los datos de el club local y visitante renombrare los datos para poder identificarlos, creando 2 nue
local=clubs
visit=clubs

for column in local.columns:
    local=local.rename(columns={column: f"{column}_local"})
    visit=visit.rename(columns={column: f"{column}_visit"})
```

In [14]:
```
#agregamos a la tabla los datos tanto del club local como el visitante
merge_table=pd.merge(merge_table.set_index("home_club_id"), local.set_index("club_id_local"), left_index=True, righ
merge_table=pd.merge(merge_table.set_index("away_club_id"), visit.set_index("club_id_visit"), left_index=True, righ
```

In [15]:
```
# vamos a hacer que la columna de fecha se reconozca de tal forma para que podamos ordenrar todo por fecha
merge_table['date'] = pd.to_datetime(merge_table['date'])
merge_table.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 122120 entries, 2 to 102261
Data columns (total 61 columns):
 #   Column                           Non-Null Count   Dtype
---  ------                           --------------   -----
 0   competition_id                   122120 non-null  object
 1   competition_type                 122120 non-null  object
 2   season                           122120 non-null  int64
 3   round                            122120 non-null  object
 4   date                             122120 non-null  datetime64[ns]
 5   home_club_goals                  122120 non-null  int64
 6   away_club_goals                  122120 non-null  int64
 7   aggregate                        122120 non-null  object
 8   home_club_position               122120 non-null  int64
 9   away_club_position               122120 non-null  int64
 10  club_home_name                   101456 non-null  object
 11  club_away_name                   103594 non-null  object
 12  home_club_manager_name           120998 non-null  object
 13  away_club_manager_name           120998 non-null  object
 14  stadium                          121752 non-null  object
 15  attendance                       122120 non-null  int64
 16  referee                          121212 non-null  object
 17  url                              122120 non-null  object
 18  club_id                          122120 non-null  int64
 19  own_goals                        122120 non-null  int64
 20  own_position                     122120 non-null  int64
 21  own_manager_name                 120998 non-null  object
 22  opponent_id                      122120 non-null  int64
 23  opponent_goals                   122120 non-null  int64
 24  opponent_position                122120 non-null  int64
 25  opponent_manager_name            120998 non-null  object
 26  hosting                          122120 non-null  object
 27  is_win                           122120 non-null  int64
 28  number_goals                     122120 non-null  int64
 29  number_game_substitutions        122120 non-null  int64
 30  goals_added_time                 122120 non-null  int64
 31  minute_last_goal                 122120 non-null  float64
 32  minute_first_goal                122120 non-null  float64
 33  club_code_local                  101456 non-null  object
 34  name_local                       101456 non-null  object
 35  domestic_competition_id_local    101456 non-null  object
```

```
 36  total_market_value_local       21624 non-null   float64
 37  squad_size_local               101456 non-null  float64
 38  average_age_local              98124 non-null   float64
 39  foreigners_number_local        101456 non-null  float64
 40  foreigners_percentage_local    96874 non-null   float64
 41  national_team_players_local    101456 non-null  float64
 42  stadium_name_local             101456 non-null  object
 43  stadium_seats_local            101456 non-null  float64
 44  net_transfer_record_local      101456 non-null  object
 45  coach_name_local               26388 non-null   object
 46  url_local                      101456 non-null  object
 47  club_code_visit                103594 non-null  object
 48  name_visit                     103594 non-null  object
 49  domestic_competition_id_visit  103594 non-null  object
 50  total_market_value_visit       22192 non-null   float64
 51  squad_size_visit               103594 non-null  float64
 52  average_age_visit              100146 non-null  float64
 53  foreigners_number_visit        103594 non-null  float64
 54  foreigners_percentage_visit    98856 non-null   float64
 55  national_team_players_visit    103594 non-null  float64
 56  stadium_name_visit             103594 non-null  object
 57  stadium_seats_visit            103594 non-null  float64
 58  net_transfer_record_visit      103594 non-null  object
 59  coach_name_visit               27096 non-null   object
 60  url_visit                      103594 non-null  object
dtypes: datetime64[ns](1), float64(16), int64(16), object(28)
memory usage: 57.8+ MB
```

In [16]:

```python
# vamos a orderar nuestro df por fecha
merge_table= merge_table.sort_values('date')
merge_table=merge_table.reset_index()
merge_table
```

Out[16]:

	away_club_id	competition_id	competition_type	season	round	date	home_club_goals	away_club_goals	aggregate	hom
0	10747	CLQ	international_cup	2011	First Round 1st leg	2012-07-03	7	0	7:0	
1	10694	ELQ	international_cup	2011	First Round 1st leg	2012-07-03	0	6	0:6	
2	5594	ELQ	international_cup	2011	First Round 1st leg	2012-07-03	0	0	0:0	
3	5594	ELQ	international_cup	2011	First Round 1st leg	2012-07-03	0	0	0:0	
4	28958	CLQ	international_cup	2011	First Round 1st leg	2012-07-03	8	0	8:0	
...	...	...	...	...	...	...	...	...	...	...
122115	868	TR1	domestic_league	2022	26. Matchday	2023-03-19	2	0	2:0	
122116	868	TR1	domestic_league	2022	26. Matchday	2023-03-19	2	0	2:0	
122117	418	ES1	domestic_league	2022	26. Matchday	2023-03-19	2	1	2:1	
122118	12	IT1	domestic_league	2022	27. Matchday	2023-03-19	1	0	1:0	
122119	1005	IT1	domestic_league	2022	27. Matchday	2023-03-19	1	0	1:0	

122120 rows × 62 columns

Ingieneria de variables

In [17]:

```
#como lo que queremos predecir son los resultados de los partidos ,vamos a desplazar las viariables que tienen que
#el resultado del partido

#para el partido pasado del local
grouped = merge_table.groupby('club_id')
merge_table['casa_goles_anteriores'] = grouped['home_club_goals'].shift()
merge_table['casa_goles_anteriores_recibidos'] = grouped['away_club_goals'].shift()
merge_table['casa_minuto_primer_gol_anterior'] = grouped['minute_first_goal'].shift()
merge_table['casa_minuto_ultimo_gol_anterior'] = grouped['minute_last_goal'].shift()
merge_table['casa_victoria_local_partidoanterior'] = grouped['is_win'].shift()
merge_table['casa_number_goals_anterior'] = grouped['number_goals'].shift()
merge_table['casa_sustituciones_anterior'] = grouped['number_game_substitutions'].shift()
merge_table['casa_goles_tiempo_añadido_anterior'] = grouped['goals_added_time'].shift()

#para el partido pasados del visitante
grouped = merge_table.groupby('away_club_id')
merge_table['visita_goles_anteriores_recibidos'] = grouped['home_club_goals'].shift()
merge_table['visita_goles_anteriores'] = grouped['away_club_goals'].shift()
merge_table['visita_minuto_primer_gol_anterior'] = grouped['minute_first_goal'].shift()
merge_table['visita_minuto_ultimo_gol_anterior'] = grouped['minute_last_goal'].shift()
merge_table['visita_victoria_local_partidoanterior'] = grouped['is_win'].shift()*-1
merge_table['visita_number_goals_anterior'] = grouped['number_goals'].shift()
merge_table['visita_sustituciones_anterior'] = grouped['number_game_substitutions'].shift()
merge_table['visita_goles_tiempo_añadido_anterior'] = grouped['goals_added_time'].shift()
merge_table.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 122120 entries, 0 to 122119
Data columns (total 78 columns):
 #   Column                                Non-Null Count   Dtype
---  ------                                --------------   -----
 0   away_club_id                          122120 non-null  int64
 1   competition_id                        122120 non-null  object
 2   competition_type                      122120 non-null  object
 3   season                                122120 non-null  int64
 4   round                                 122120 non-null  object
 5   date                                  122120 non-null  datetime64[ns]
 6   home_club_goals                       122120 non-null  int64
 7   away_club_goals                       122120 non-null  int64
 8   aggregate                             122120 non-null  object
 9   home_club_position                    122120 non-null  int64
 10  away_club_position                    122120 non-null  int64
 11  club_home_name                        101456 non-null  object
 12  club_away_name                        103594 non-null  object
 13  home_club_manager_name                120998 non-null  object
 14  away_club_manager_name                120998 non-null  object
 15  stadium                               121752 non-null  object
 16  attendance                            122120 non-null  int64
 17  referee                               121212 non-null  object
 18  url                                   122120 non-null  object
 19  club_id                               122120 non-null  int64
 20  own_goals                             122120 non-null  int64
 21  own_position                          122120 non-null  int64
 22  own_manager_name                      120998 non-null  object
 23  opponent_id                           122120 non-null  int64
 24  opponent_goals                        122120 non-null  int64
 25  opponent_position                     122120 non-null  int64
 26  opponent_manager_name                 120998 non-null  object
 27  hosting                               122120 non-null  object
 28  is_win                                122120 non-null  int64
 29  number_goals                          122120 non-null  int64
 30  number_game_substitutions             122120 non-null  int64
 31  goals_added_time                      122120 non-null  int64
 32  minute_last_goal                      122120 non-null  float64
 33  minute_first_goal                     122120 non-null  float64
 34  club_code_local                       101456 non-null  object
 35  name_local                            101456 non-null  object
```

```
36  domestic_competition_id_local          101456 non-null  object
37  total_market_value_local               21624 non-null   float64
38  squad_size_local                       101456 non-null  float64
39  average_age_local                      98124 non-null   float64
40  foreigners_number_local                101456 non-null  float64
41  foreigners_percentage_local            96874 non-null   float64
42  national_team_players_local            101456 non-null  float64
43  stadium_name_local                     101456 non-null  object
44  stadium_seats_local                    101456 non-null  float64
45  net_transfer_record_local              101456 non-null  object
46  coach_name_local                       26388 non-null   object
47  url_local                              101456 non-null  object
48  club_code_visit                        103594 non-null  object
49  name_visit                             103594 non-null  object
50  domestic_competition_id_visit          103594 non-null  object
51  total_market_value_visit               22192 non-null   float64
52  squad_size_visit                       103594 non-null  float64
53  average_age_visit                      100146 non-null  float64
54  foreigners_number_visit                103594 non-null  float64
55  foreigners_percentage_visit            98856 non-null   float64
56  national_team_players_visit            103594 non-null  float64
57  stadium_name_visit                     103594 non-null  object
58  stadium_seats_visit                    103594 non-null  float64
59  net_transfer_record_visit              103594 non-null  object
60  coach_name_visit                       27096 non-null   object
61  url_visit                              103594 non-null  object
62  casa_goles_anteriores                  119568 non-null  float64
63  casa_goles_anteriores_recibidos        119568 non-null  float64
64  casa_minuto_primer_gol_anterior        119568 non-null  float64
65  casa_minuto_ultimo_gol_anterior        119568 non-null  float64
66  casa_victoria_local_partidoanterior    119568 non-null  float64
67  casa_number_goals_anterior             119568 non-null  float64
68  casa_sustituciones_anterior            119568 non-null  float64
69  casa_goles_tiempo_añadido_anterior     119568 non-null  float64
70  visita_goles_anteriores_recibidos      120080 non-null  float64
71  visita_goles_anteriores                120080 non-null  float64
72  visita_minuto_primer_gol_anterior      120080 non-null  float64
73  visita_minuto_ultimo_gol_anterior      120080 non-null  float64
74  visita_victoria_local_partidoanterior  120080 non-null  float64
75  visita_number_goals_anterior           120080 non-null  float64
76  visita_sustituciones_anterior          120080 non-null  float64
```

```
 77  visita_goles_tiempo_añadido_anterior   120080 non-null  float64
dtypes: datetime64[ns](1), float64(32), int64(17), object(28)
memory usage: 72.7+ MB
```

In [18]:

```python
#como lo que queremos predecir son los resultados de los partidos en realida,vamos a desplazar las viariables que t
#el resultado del partido ahora dando info de los 2 partidos previos

#para el partido pasado del local
grouped = merge_table.groupby('club_id')
merge_table['casa_goles_anteriores_2'] = grouped['home_club_goals'].shift(2)
merge_table['casa_goles_anteriores_recibidos_2'] = grouped['away_club_goals'].shift(2)
merge_table['casa_minuto_primer_gol_anterior_2'] = grouped['minute_first_goal'].shift(2)
merge_table['casa_minuto_ultimo_gol_anterior_2'] = grouped['minute_last_goal'].shift(2)
merge_table['casa_victoria_local_partidoanterior_2'] = grouped['is_win'].shift(2)
merge_table['casa_number_goals_anterior_2'] = grouped['number_goals'].shift(2)
merge_table['casa_sustituciones_anterior_2'] = grouped['number_game_substitutions'].shift(2)
merge_table['casa_goles_tiempo_añadido_anterior_2'] = grouped['goals_added_time'].shift(2)

#para el partido pasados del visitante
grouped = merge_table.groupby('away_club_id')
merge_table['visita_goles_anteriores_recibidos_2'] = grouped['home_club_goals'].shift(2)
merge_table['visita_goles_anteriores_2'] = grouped['away_club_goals'].shift(2)
merge_table['visita_minuto_primer_gol_anterior_2'] = grouped['minute_first_goal'].shift(2)
merge_table['visita_minuto_ultimo_gol_anterior_2'] = grouped['minute_last_goal'].shift(2)
merge_table['visita_victoria_local_partidoanterior_2'] = grouped['is_win'].shift(2)*-1
merge_table['visita_number_goals_anterior_2'] = grouped['number_goals'].shift(2)
merge_table['visita_sustituciones_anterior_2'] = grouped['number_game_substitutions'].shift(2)
merge_table['visita_goles_tiempo_añadido_anterior_2'] = grouped['goals_added_time'].shift(2)
merge_table.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 122120 entries, 0 to 122119
Data columns (total 94 columns):
 #   Column                                Non-Null Count   Dtype
---  ------                                --------------   -----
 0   away_club_id                          122120 non-null  int64
 1   competition_id                        122120 non-null  object
 2   competition_type                      122120 non-null  object
 3   season                                122120 non-null  int64
 4   round                                 122120 non-null  object
 5   date                                  122120 non-null  datetime64[ns]
 6   home_club_goals                       122120 non-null  int64
 7   away_club_goals                       122120 non-null  int64
 8   aggregate                             122120 non-null  object
 9   home_club_position                    122120 non-null  int64
 10  away_club_position                    122120 non-null  int64
 11  club_home_name                        101456 non-null  object
 12  club_away_name                        103594 non-null  object
 13  home_club_manager_name                120998 non-null  object
 14  away_club_manager_name                120998 non-null  object
 15  stadium                               121752 non-null  object
 16  attendance                            122120 non-null  int64
 17  referee                               121212 non-null  object
 18  url                                   122120 non-null  object
 19  club_id                               122120 non-null  int64
 20  own_goals                             122120 non-null  int64
 21  own_position                          122120 non-null  int64
 22  own_manager_name                      120998 non-null  object
 23  opponent_id                           122120 non-null  int64
 24  opponent_goals                        122120 non-null  int64
 25  opponent_position                     122120 non-null  int64
 26  opponent_manager_name                 120998 non-null  object
 27  hosting                               122120 non-null  object
 28  is_win                                122120 non-null  int64
 29  number_goals                          122120 non-null  int64
 30  number_game_substitutions             122120 non-null  int64
 31  goals_added_time                      122120 non-null  int64
 32  minute_last_goal                      122120 non-null  float64
 33  minute_first_goal                     122120 non-null  float64
 34  club_code_local                       101456 non-null  object
 35  name_local                            101456 non-null  object
```

```
36  domestic_competition_id_local           101456 non-null  object
37  total_market_value_local                21624 non-null   float64
38  squad_size_local                        101456 non-null  float64
39  average_age_local                       98124 non-null   float64
40  foreigners_number_local                 101456 non-null  float64
41  foreigners_percentage_local             96874 non-null   float64
42  national_team_players_local             101456 non-null  float64
43  stadium_name_local                      101456 non-null  object
44  stadium_seats_local                     101456 non-null  float64
45  net_transfer_record_local               101456 non-null  object
46  coach_name_local                        26388 non-null   object
47  url_local                               101456 non-null  object
48  club_code_visit                         103594 non-null  object
49  name_visit                              103594 non-null  object
50  domestic_competition_id_visit           103594 non-null  object
51  total_market_value_visit                22192 non-null   float64
52  squad_size_visit                        103594 non-null  float64
53  average_age_visit                       100146 non-null  float64
54  foreigners_number_visit                 103594 non-null  float64
55  foreigners_percentage_visit             98856 non-null   float64
56  national_team_players_visit             103594 non-null  float64
57  stadium_name_visit                      103594 non-null  object
58  stadium_seats_visit                     103594 non-null  float64
59  net_transfer_record_visit               103594 non-null  object
60  coach_name_visit                        27096 non-null   object
61  url_visit                               103594 non-null  object
62  casa_goles_anteriores                   119568 non-null  float64
63  casa_goles_anteriores_recibidos         119568 non-null  float64
64  casa_minuto_primer_gol_anterior         119568 non-null  float64
65  casa_minuto_ultimo_gol_anterior         119568 non-null  float64
66  casa_victoria_local_partidoanterior     119568 non-null  float64
67  casa_number_goals_anterior              119568 non-null  float64
68  casa_sustituciones_anterior             119568 non-null  float64
69  casa_goles_tiempo_añadido_anterior      119568 non-null  float64
70  visita_goles_anteriores_recibidos       120080 non-null  float64
71  visita_goles_anteriores                 120080 non-null  float64
72  visita_minuto_primer_gol_anterior       120080 non-null  float64
73  visita_minuto_ultimo_gol_anterior       120080 non-null  float64
74  visita_victoria_local_partidoanterior   120080 non-null  float64
75  visita_number_goals_anterior            120080 non-null  float64
76  visita_sustituciones_anterior           120080 non-null  float64
```

```
 77  visita_goles_tiempo_añadido_anterior     120080 non-null  float64
 78  casa_goles_anteriores_2                  117529 non-null  float64
 79  casa_goles_anteriores_recibidos_2        117529 non-null  float64
 80  casa_minuto_primer_gol_anterior_2        117529 non-null  float64
 81  casa_minuto_ultimo_gol_anterior_2        117529 non-null  float64
 82  casa_victoria_local_partidoanterior_2    117529 non-null  float64
 83  casa_number_goals_anterior_2             117529 non-null  float64
 84  casa_sustituciones_anterior_2            117529 non-null  float64
 85  casa_goles_tiempo_añadido_anterior_2     117529 non-null  float64
 86  visita_goles_anteriores_recibidos_2      118040 non-null  float64
 87  visita_goles_anteriores_2                118040 non-null  float64
 88  visita_minuto_primer_gol_anterior_2      118040 non-null  float64
 89  visita_minuto_ultimo_gol_anterior_2      118040 non-null  float64
 90  visita_victoria_local_partidoanterior_2  118040 non-null  float64
 91  visita_number_goals_anterior_2           118040 non-null  float64
 92  visita_sustituciones_anterior_2          118040 non-null  float64
 93  visita_goles_tiempo_añadido_anterior_2   118040 non-null  float64
dtypes: datetime64[ns](1), float64(48), int64(17), object(28)
memory usage: 87.6+ MB
```

In [19]:

```python
#como lo que queremos predecir son los resultados de los partidos en realida,vamos a desplazar las viariables que t
#el resultado del partido ahora dando info de los 3 partidos previos

#para el partido pasado del local
grouped = merge_table.groupby('club_id')
merge_table['casa_goles_anteriores_3'] = grouped['home_club_goals'].shift(3)
merge_table['casa_goles_anteriores_recibidos_3'] = grouped['away_club_goals'].shift(3)
merge_table['casa_minuto_primer_gol_anterior_3'] = grouped['minute_first_goal'].shift(3)
merge_table['casa_minuto_ultimo_gol_anterior_3'] = grouped['minute_last_goal'].shift(3)
merge_table['casa_victoria_local_partidoanterior_3'] = grouped['is_win'].shift(3)
merge_table['casa_number_goals_anterior_3'] = grouped['number_goals'].shift(3)
merge_table['casa_sustituciones_anterior_3'] = grouped['number_game_substitutions'].shift(3)
merge_table['casa_goles_tiempo_añadido_anterior_3'] = grouped['goals_added_time'].shift(3)

#para el partido pasados del visitante
grouped = merge_table.groupby('away_club_id')
merge_table['visita_goles_anteriores_recibidos_3'] = grouped['home_club_goals'].shift(3)
merge_table['visita_goles_anteriores_3'] = grouped['away_club_goals'].shift(3)
merge_table['visita_minuto_primer_gol_anterior_3'] = grouped['minute_first_goal'].shift(3)
merge_table['visita_minuto_ultimo_gol_anterior_3'] = grouped['minute_last_goal'].shift(3)
merge_table['visita_victoria_local_partidoanterior_3'] = grouped['is_win'].shift(3)*-1
```

```
merge_table['visita_number_goals_anterior_3'] = grouped['number_goals'].shift(3)
merge_table['visita_sustituciones_anterior_3'] = grouped['number_game_substitutions'].shift(3)
merge_table['visita_goles_tiempo_añadido_anterior_3'] = grouped['goals_added_time'].shift(3)
merge_table.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 122120 entries, 0 to 122119
Columns: 110 entries, away_club_id to visita_goles_tiempo_añadido_anterior_3
dtypes: datetime64[ns](1), float64(64), int64(17), object(28)
memory usage: 102.5+ MB
```

In [20]:

```
# siguiendo el mismo pensamiento de que queremos predecir el total de goles, antes de un partido
#eleminaremos las variables correspondientes alos sucesos ocurridos durante el partido, ya que no tiene sentido
#querer predecir total de goles, si el partido ya termino, a excepcion de nuestra variable objetivo
#tambien eliminaremos variables que no tienen sentido, como tamaño de estadio de la visita, ya que es un factor tot
#y variables que venian repetidas por dos tablas
ls_delte=['stadium_seats_visit',"home_club_goals","away_club_goals","minute_first_goal","minute_last_goal",'is_win'

for i in ls_delte:
    merge_table=merge_table.drop([i],axis=1)

merge_table.sample(5)
```

Out[20]:

	away_club_id	competition_id	competition_type	season	round	date	home_club_position	away_club_position	club_home_
45159	2700	RU1	domestic_league	2015	26. Matchday	2016-05-01	3	16	Fk Kras
544	2227	UKR1	domestic_league	2011	3. Matchday	2012-07-28	8	14	Fk Ma
88139	1005	IT1	domestic_league	2019	25. Matchday	2020-02-23	5	16	As
92456	16239	UKR1	domestic_league	2020	6. Matchday	2020-10-18	10	7	Sk Dn
79645	38204	ELQ	international_cup	2018	Second Round 1st leg	2019-07-23	-1	-1	

5 rows × 96 columns

In [21]:
```
merge_table=merge_table.sort_values('date')
```

In [23]:
```
# creacion variable objetivo
merge_table["over_2"]=merge_table["number_goals"].map(lambda x: 1 if x > 2 else 0)
```

In [24]:
```
# Veamos cuales son las variables continuas para crear mas variables a partir de ellas
ls_cont=[]
for i in  merge_table.columns:
        if not  merge_table[i].dtype == object:
            if merge_table[i].nunique() > 2:
                ls_cont.append(i)
```

In [25]:
```
merge_table[ls_cont].columns
```

Out[25]:
```
Index(['away_club_id', 'season', 'date', 'home_club_position',
       'away_club_position', 'attendance', 'club_id', 'own_position',
       'opponent_position', 'number_goals', 'total_market_value_local',
       'squad_size_local', 'average_age_local', 'foreigners_number_local',
       'foreigners_percentage_local', 'national_team_players_local',
       'stadium_seats_local', 'total_market_value_visit', 'squad_size_visit',
       'average_age_visit', 'foreigners_number_visit',
       'foreigners_percentage_visit', 'national_team_players_visit',
       'casa_goles_anteriores', 'casa_goles_anteriores_recibidos',
       'casa_minuto_primer_gol_anterior', 'casa_minuto_ultimo_gol_anterior',
       'casa_number_goals_anterior', 'casa_sustituciones_anterior',
       'casa_goles_tiempo_añadido_anterior',
       'visita_goles_anteriores_recibidos', 'visita_goles_anteriores',
       'visita_minuto_primer_gol_anterior',
       'visita_minuto_ultimo_gol_anterior', 'visita_number_goals_anterior',
       'visita_sustituciones_anterior', 'visita_goles_tiempo_añadido_anterior',
       'casa_goles_anteriores_2', 'casa_goles_anteriores_recibidos_2',
       'casa_minuto_primer_gol_anterior_2',
       'casa_minuto_ultimo_gol_anterior_2', 'casa_number_goals_anterior_2',
       'casa_sustituciones_anterior_2', 'casa_goles_tiempo_añadido_anterior_2',
       'visita_goles_anteriores_recibidos_2', 'visita_goles_anteriores_2',
       'visita_minuto_primer_gol_anterior_2',
       'visita_minuto_ultimo_gol_anterior_2', 'visita_number_goals_anterior_2',
       'visita_sustituciones_anterior_2',
       'visita_goles_tiempo_añadido_anterior_2', 'casa_goles_anteriores_3',
       'casa_goles_anteriores_recibidos_3',
       'casa_minuto_primer_gol_anterior_3',
       'casa_minuto_ultimo_gol_anterior_3', 'casa_number_goals_anterior_3',
       'casa_sustituciones_anterior_3', 'casa_goles_tiempo_añadido_anterior_3',
       'visita_goles_anteriores_recibidos_3', 'visita_goles_anteriores_3',
       'visita_minuto_primer_gol_anterior_3',
       'visita_minuto_ultimo_gol_anterior_3', 'visita_number_goals_anterior_3',
       'visita_sustituciones_anterior_3',
       'visita_goles_tiempo_añadido_anterior_3'],
      dtype='object')
```

In [26]:
```
# Calcular el promedio actualizado de goles recibidos por equipo utilizando rolling y groupby

merge_table["casa_goles_recibidos_mean"] = merge_table.groupby('club_id')['casa_goles_anteriores_recibidos'].rollin
merge_table["casa_goles_anotados_mean"]= merge_table.groupby('club_id')['casa_goles_anteriores'].rolling(window=5,
```

```
merge_table["visita_goles_recibidos_mean"] = merge_table.groupby('away_club_id')['visita_goles_anteriores_recibidos
merge_table["visita_goles_anotados_mean"]= merge_table.groupby('away_club_id')['visita_goles_anteriores'].rolling(w
merge_table["casa_number_goals_mean"]= merge_table.groupby('away_club_id')['casa_number_goals_anterior'].rolling(wi
merge_table["visita_number_goals_mean"]= merge_table.groupby('away_club_id')['visita_number_goals_anterior'].rollin
```

In [27]:
```
# Creacion de variables continuas para 1 partido anterior
merge_table["age_difference"]=merge_table['average_age_local']-merge_table['average_age_visit']
merge_table["foreigners_percentage_difference"]=merge_table['foreigners_percentage_local']-merge_table['foreigners_
merge_table["suma_goles_recibidos"]=merge_table['casa_goles_anteriores_recibidos']+merge_table['visita_goles_anteri
merge_table["suma_goles_anotados"]=merge_table['casa_goles_anteriores']+merge_table['visita_goles_anteriores']
merge_table["total_de_goles_anterior_casa"]=merge_table['casa_goles_anteriores']+merge_table['casa_goles_anteriores
merge_table["total_de_goles_anterior_visita"]=merge_table['visita_goles_anteriores']+merge_table['visita_goles_ante
merge_table["anotado_local_recibido_visita"]=merge_table['casa_goles_anteriores']+merge_table['visita_goles_anterio
merge_table["anotado_visita_recibido_local"]=merge_table['casa_goles_anteriores_recibidos']+merge_table['visita_gol
merge_table['over_2_anterior_casa']=merge_table["total_de_goles_anterior_casa"].map(lambda x: 1 if x > 2 else 0)
merge_table['over_2_anterior_visita']=merge_table["total_de_goles_anterior_visita"].map(lambda x: 1 if x > 2 else 0
merge_table["diferencia_promedio_anotados"]=merge_table["casa_goles_anotados_mean"]-merge_table["visita_goles_anota
merge_table["diferencia_promedio_recibidos"]=merge_table["casa_goles_recibidos_mean"]-merge_table["visita_goles_rec
merge_table["suma_promedio_recibidos_anotados"]=merge_table["casa_goles_recibidos_mean"]+merge_table["visita_goles_
merge_table["suma_promedio_anotados_recibidos"]=merge_table["casa_goles_anotados_mean"]+merge_table["visita_goles_r
merge_table["suma_promedio_number_goals"]=merge_table["casa_number_goals_mean"]+merge_table["visita_number_goals_me
merge_table["diferencia_promedio_number_goals"]=merge_table["casa_number_goals_mean"]-merge_table["visita_number_go
merge_table["diferencia_sustituciones"]=merge_table['casa_sustituciones_anterior']-merge_table['visita_sustitucione
```

In [28]:
```
# Creacion de variables continuas para 2 partidos anteriores
merge_table["suma_goles_recibidos_2"]=merge_table['casa_goles_anteriores_recibidos_2']+merge_table['visita_goles_an
merge_table["suma_goles_anotados_2"]=merge_table['casa_goles_anteriores_2']+merge_table['visita_goles_anteriores_2'
merge_table["total_de_goles_anterior_casa_2"]=merge_table['casa_goles_anteriores_2']+merge_table['casa_goles_anteri
merge_table["total_de_goles_anterior_visita_2"]=merge_table['visita_goles_anteriores_2']+merge_table['visita_goles_
merge_table["anotado_local_recibido_visita_2"]=merge_table['casa_goles_anteriores_2']+merge_table['visita_goles_ant
merge_table["anotado_visita_recibido_local_2"]=merge_table['casa_goles_anteriores_recibidos_2']+merge_table['visita
merge_table['over_2_anterior_casa_2']=merge_table["total_de_goles_anterior_casa_2"].map(lambda x: 1 if x > 2 else 0
merge_table['over_2_anterior_visita_2']=merge_table["total_de_goles_anterior_visita_2"].map(lambda x: 1 if x > 2 el
merge_table["diferencia_sustituciones_2"]=merge_table['casa_sustituciones_anterior_2']-merge_table['visita_sustituc
```

In [29]:
```
# Creacion de variables continuas para 3 partidos anteriores
merge_table["suma_goles_recibidos_3"]=merge_table['casa_goles_anteriores_recibidos_3']+merge_table['visita_goles_an
merge_table["suma_goles_anotados_3"]=merge_table['casa_goles_anteriores_3']+merge_table['visita_goles_anteriores_3'
merge_table["total_de_goles_anterior_casa_3"]=merge_table['casa_goles_anteriores_3']+merge_table['casa_goles_anteri
merge_table["total_de_goles_anterior_visita_3"]=merge_table['visita_goles_anteriores_3']+merge_table['visita_goles_
```

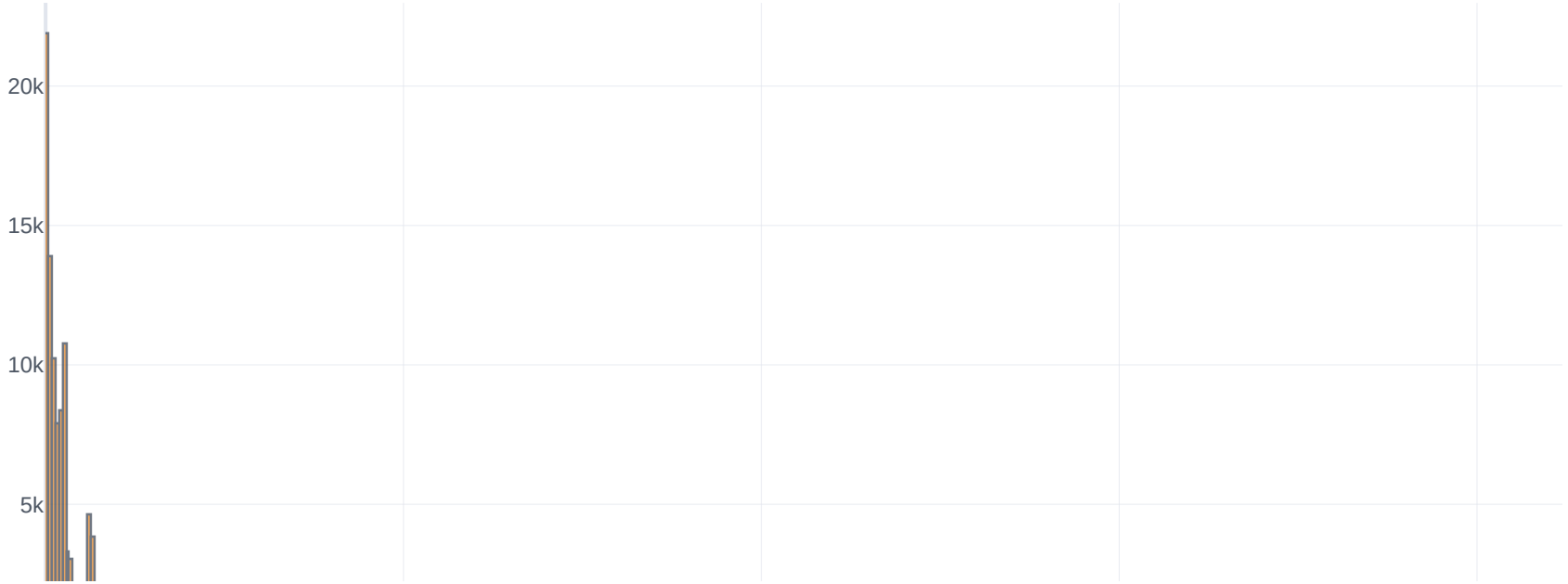
```
merge_table["anotado_local_recibido_visita_3"]=merge_table['casa_goles_anteriores_3']+merge_table['visita_goles_ant
merge_table["anotado_visita_recibido_local_3"]=merge_table['casa_goles_anteriores_recibidos_3']+merge_table['visita
merge_table['over_2_anterior_casa_3']=merge_table["total_de_goles_anterior_casa_3"].map(lambda x: 1 if x > 2 else 0
merge_table['over_2_anterior_visita_3']=merge_table["total_de_goles_anterior_visita_3"].map(lambda x: 1 if x > 2 el
merge_table["diferencia_sustituciones_3"]=merge_table['casa_sustituciones_anterior_3']-merge_table['visita_sustituc
```

In [30]:
```
merge_table=merge_table.set_index('date')
```

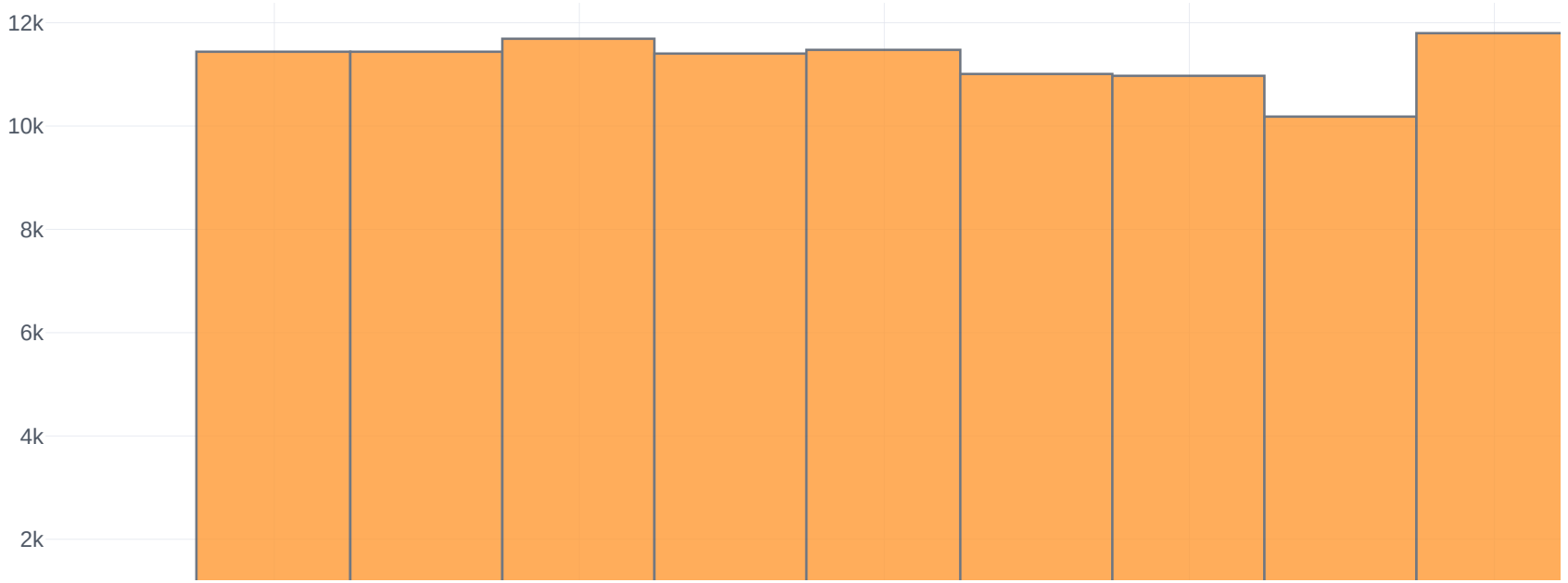
EDA

In [31]:
```
# Distribucion de variables continuas
ls_cont=[]
for i in  merge_table.columns:
    if not  merge_table[i].dtype == object:
        if merge_table[i].nunique() > 2:
            res=merge_table[i].map(lambda x:float(x)).iplot(kind="hist", title=f"{i}'s histogram", asFigure=True)
            res.show()#descomentar para eda
            ls_cont.append(i)
```

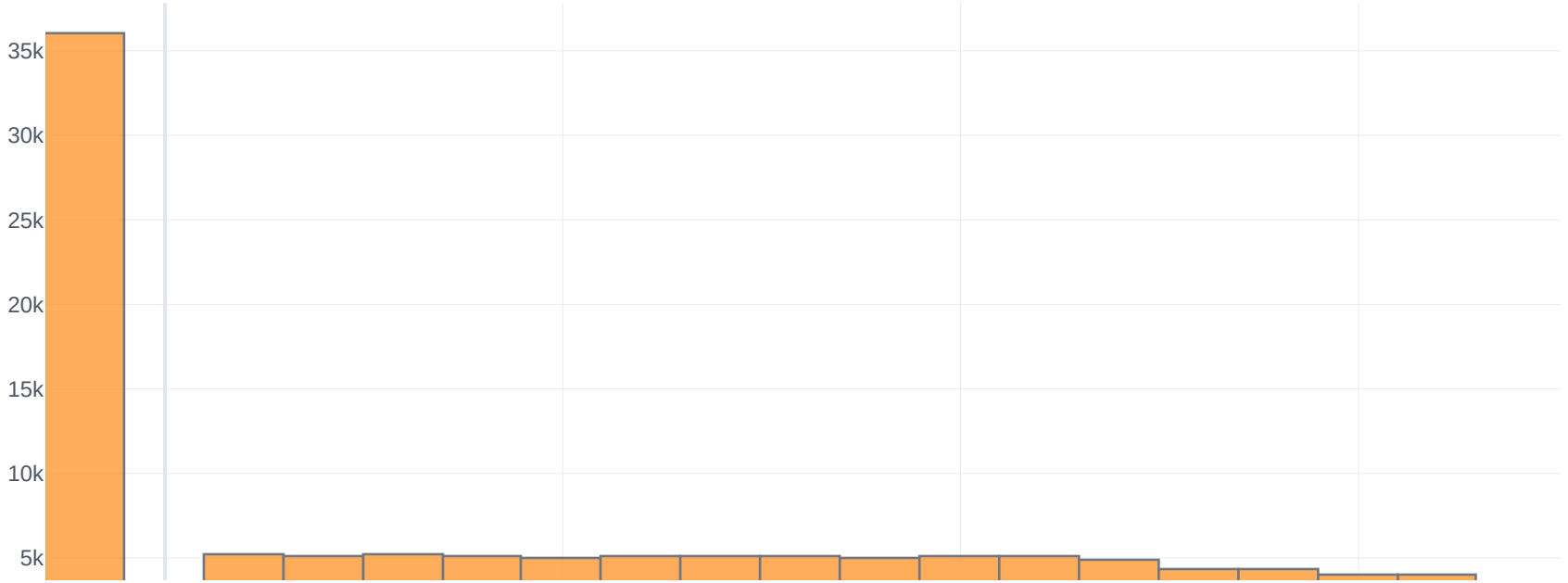
away_club_id's histogram



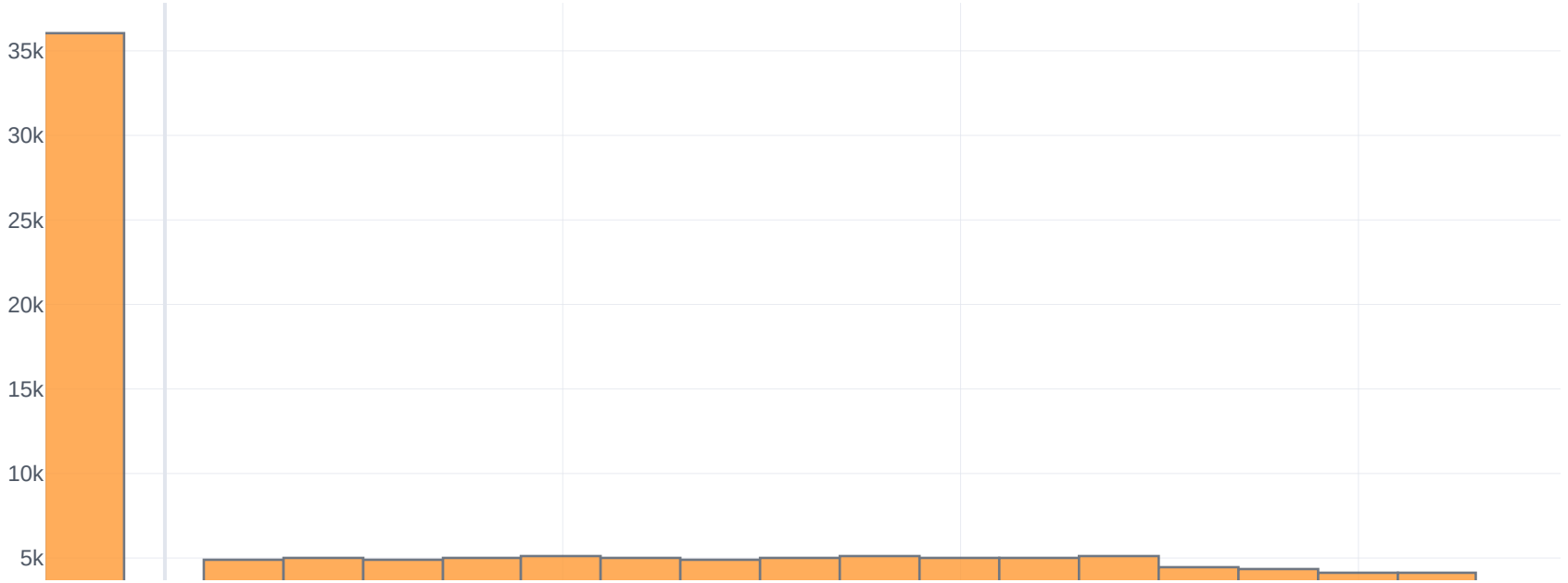
season's histogram



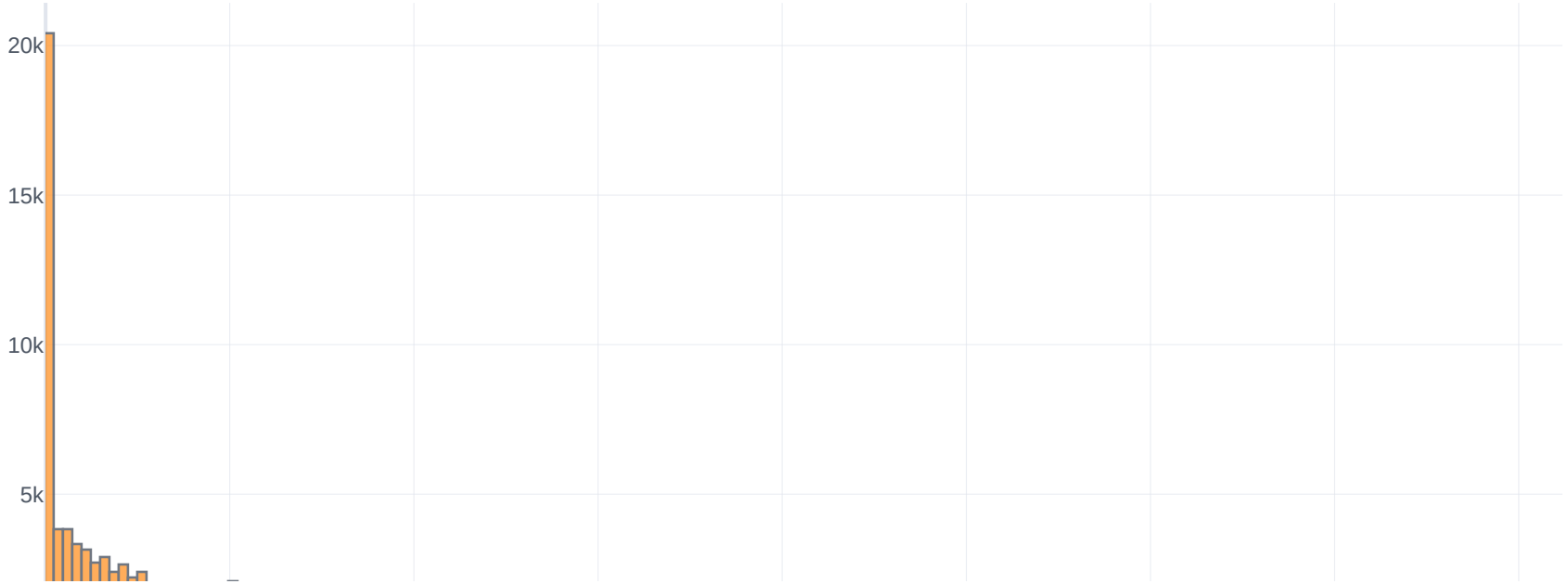
home_club_position's histogram



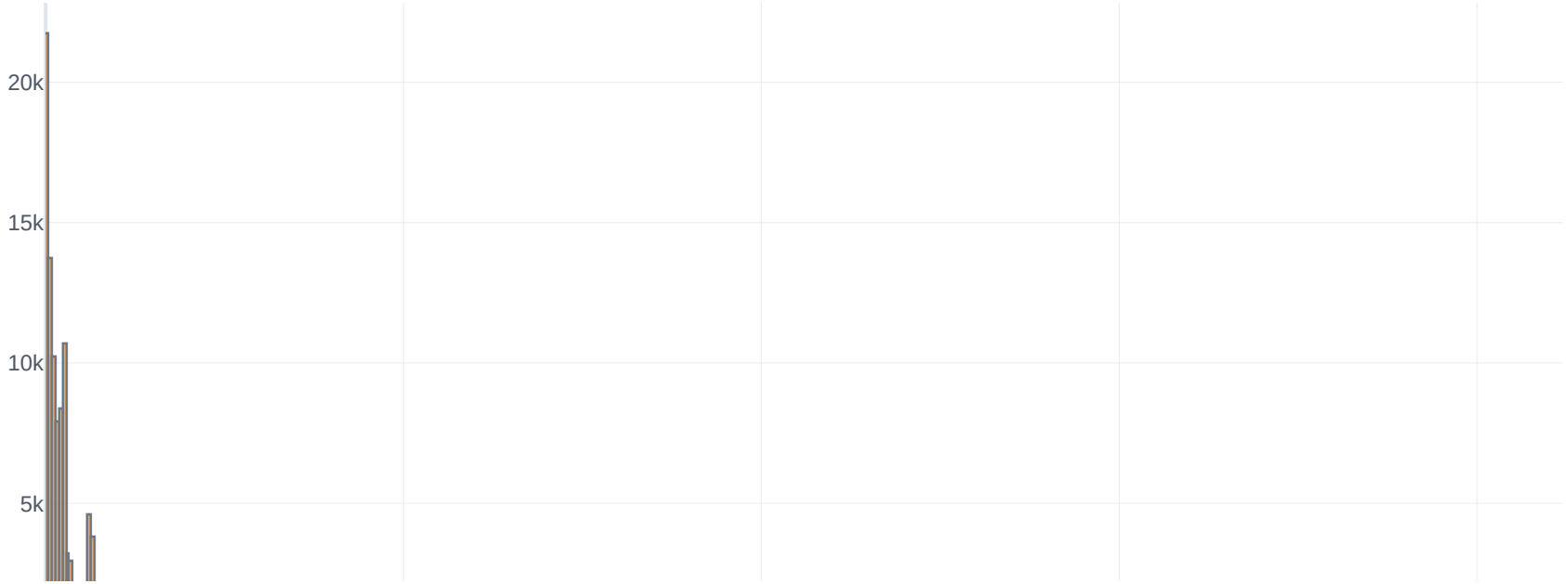
away_club_position's histogram



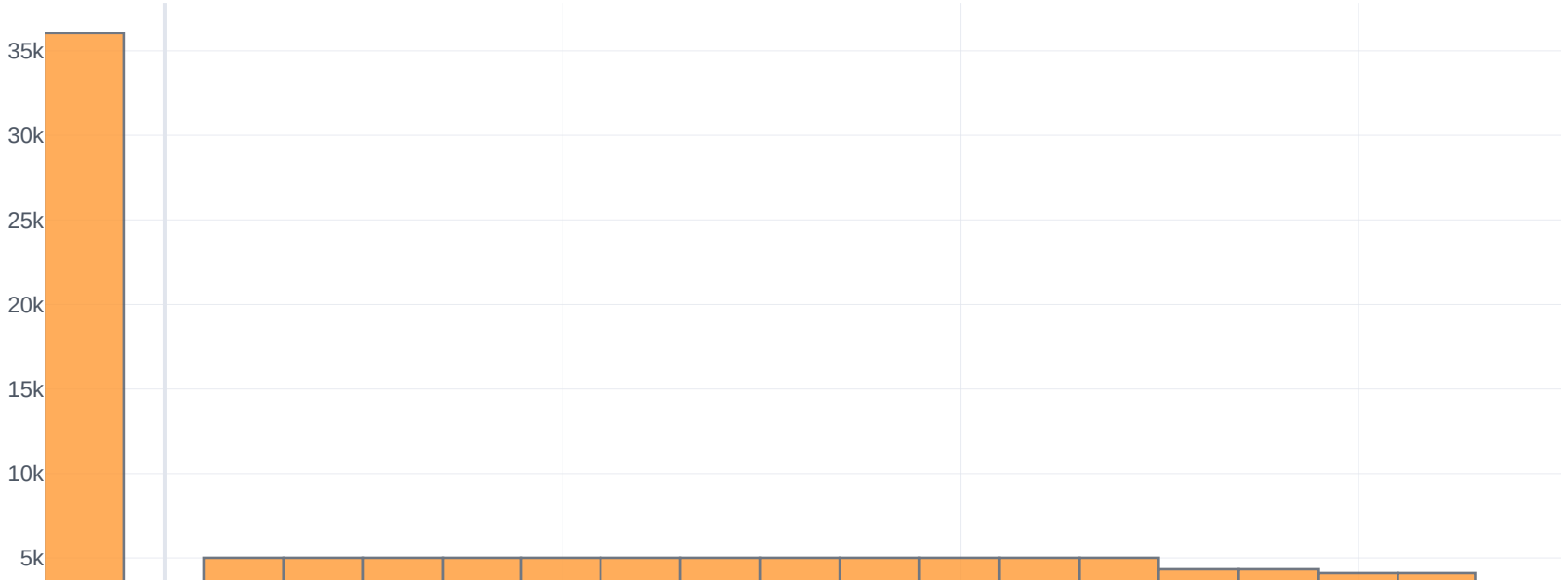
attendance's histogram



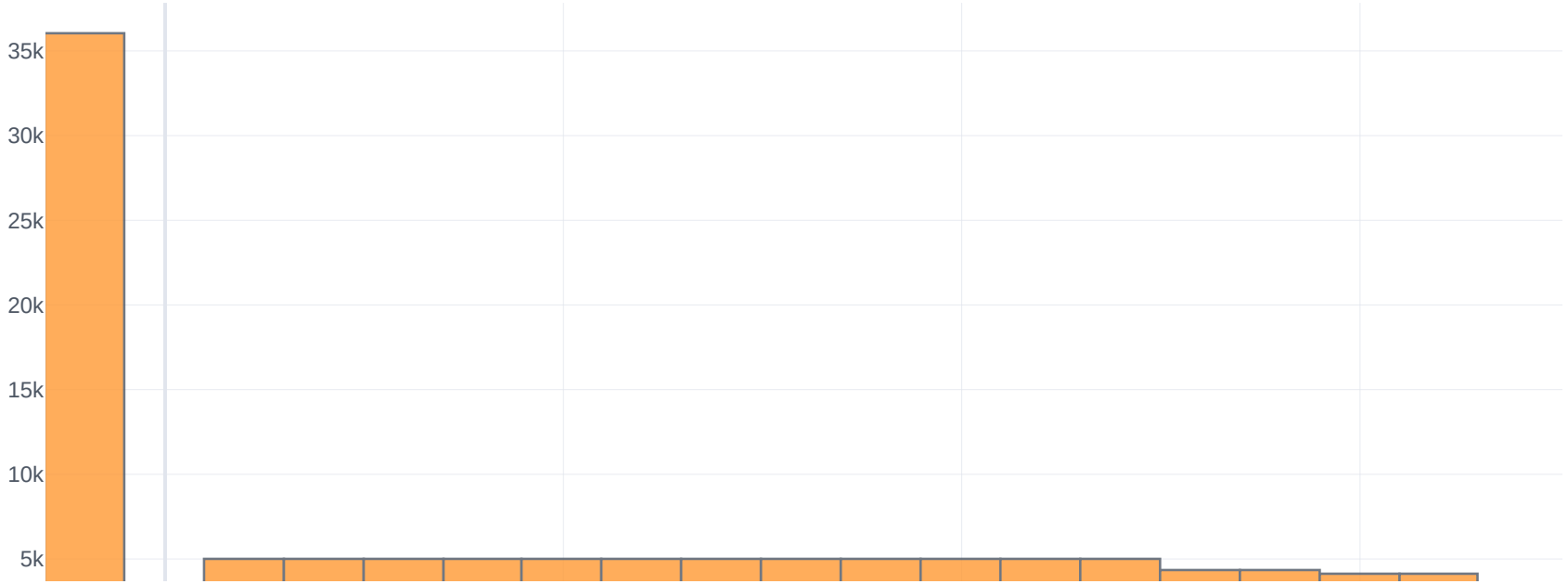
club_id's histogram



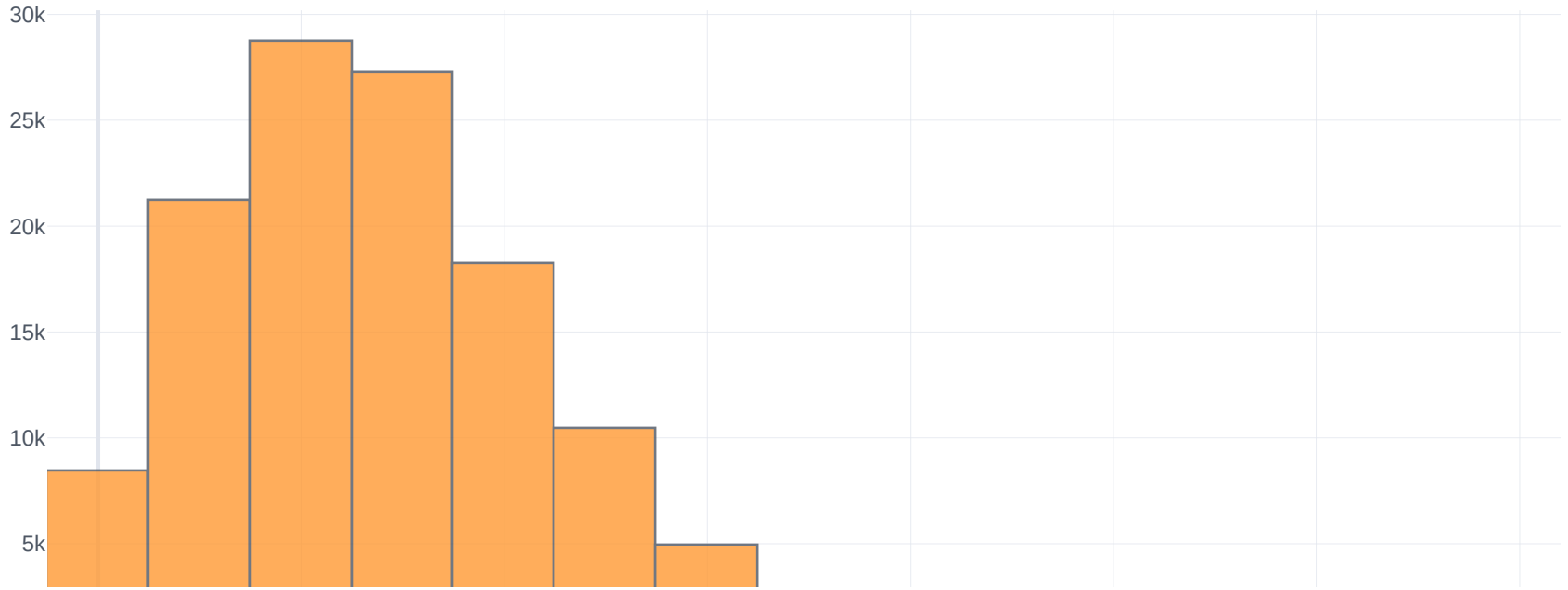
own_position's histogram



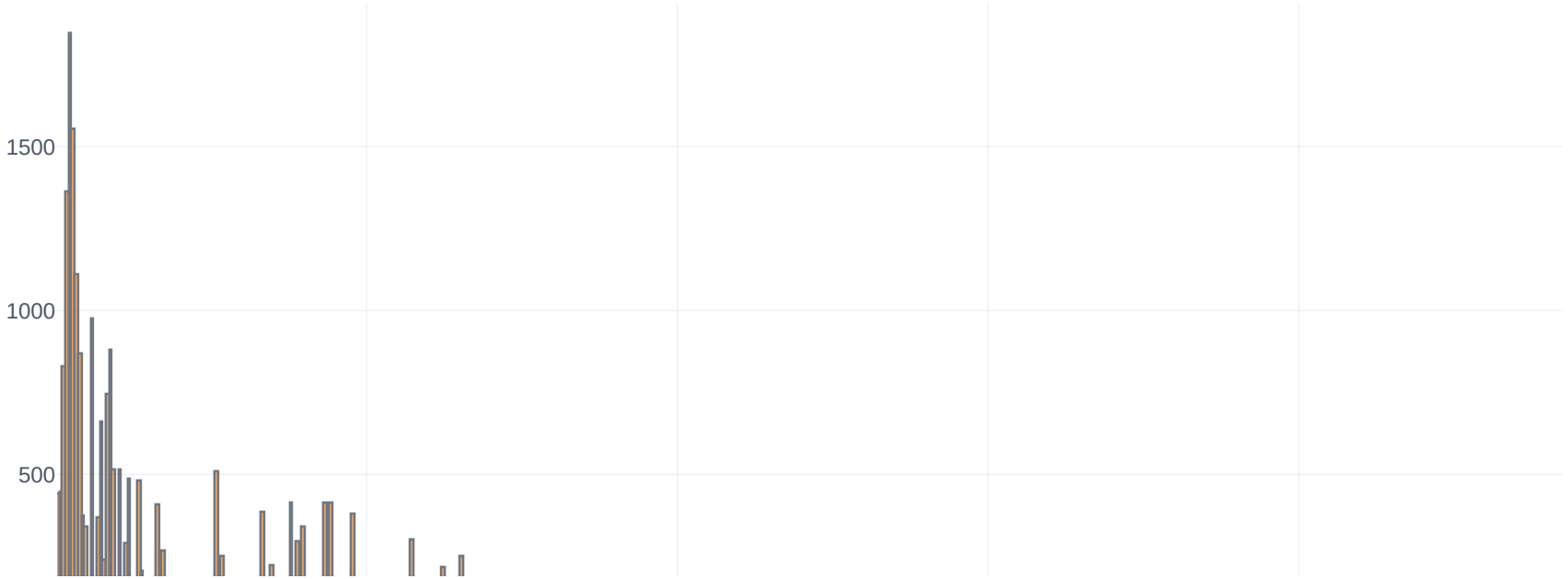
opponent_position's histogram

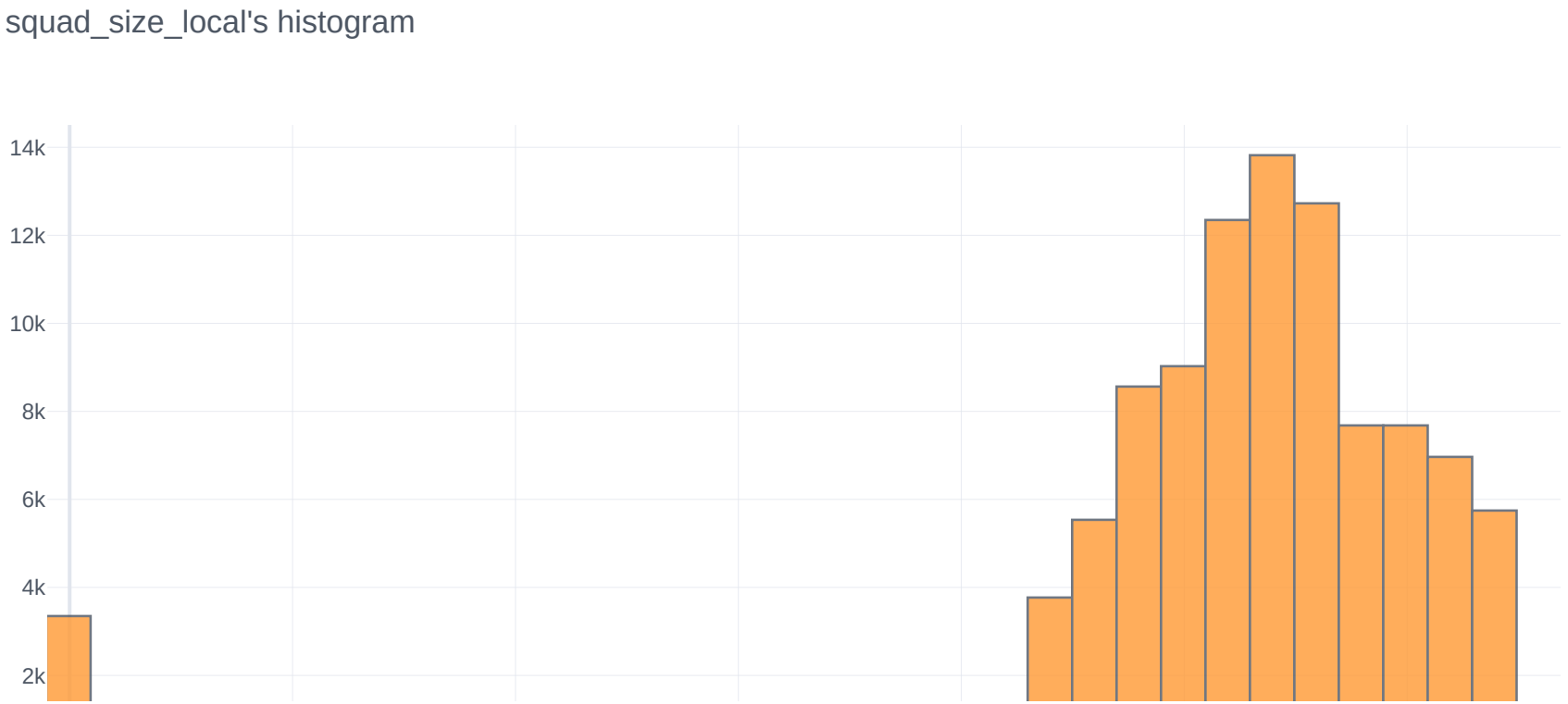


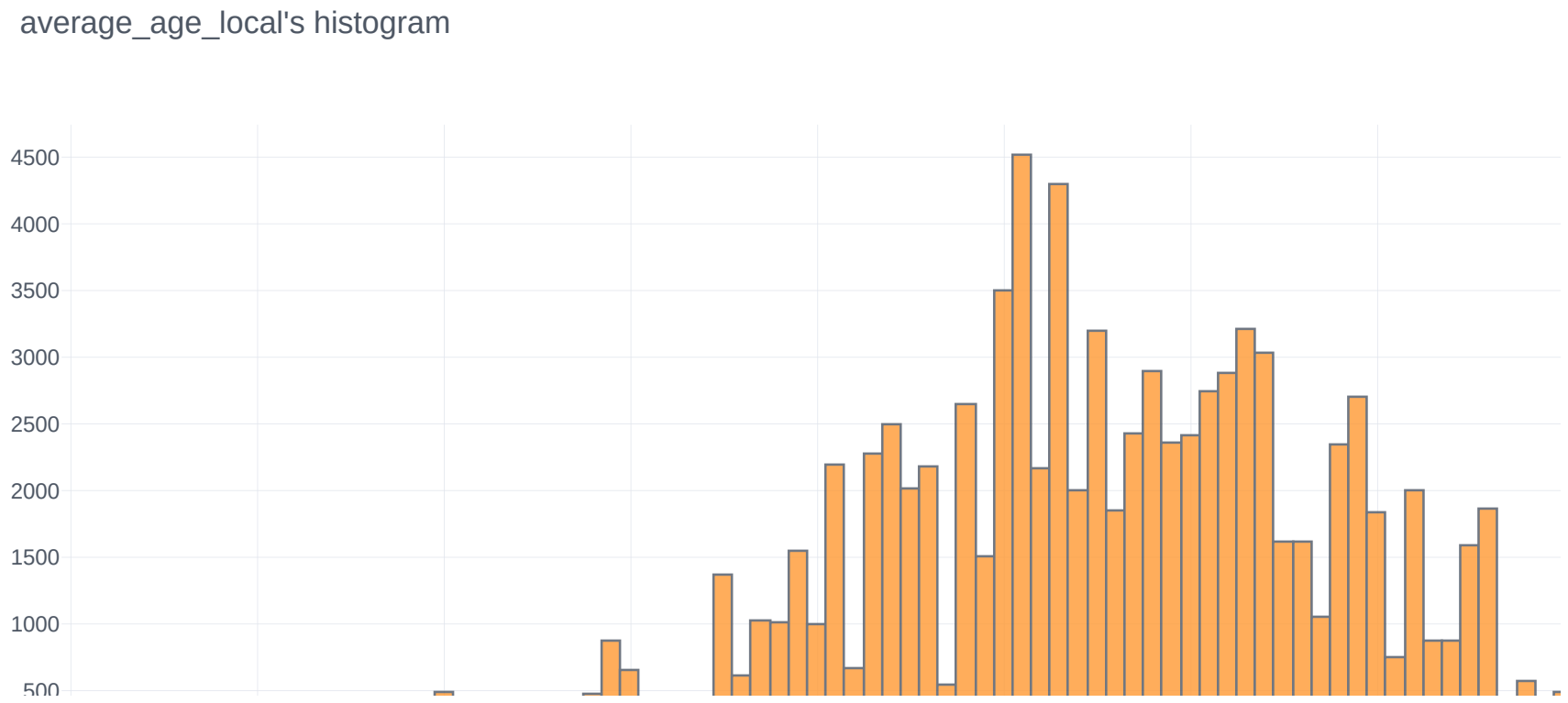
number_goals's histogram



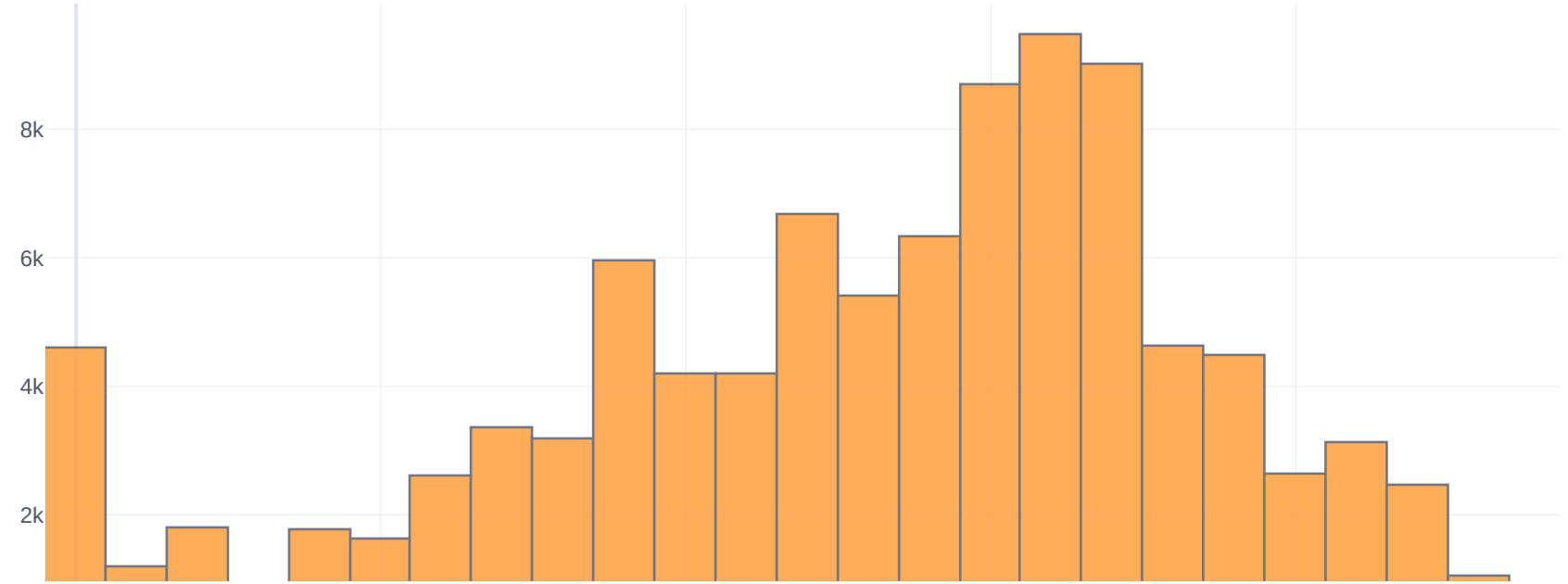
total_market_value_local's histogram



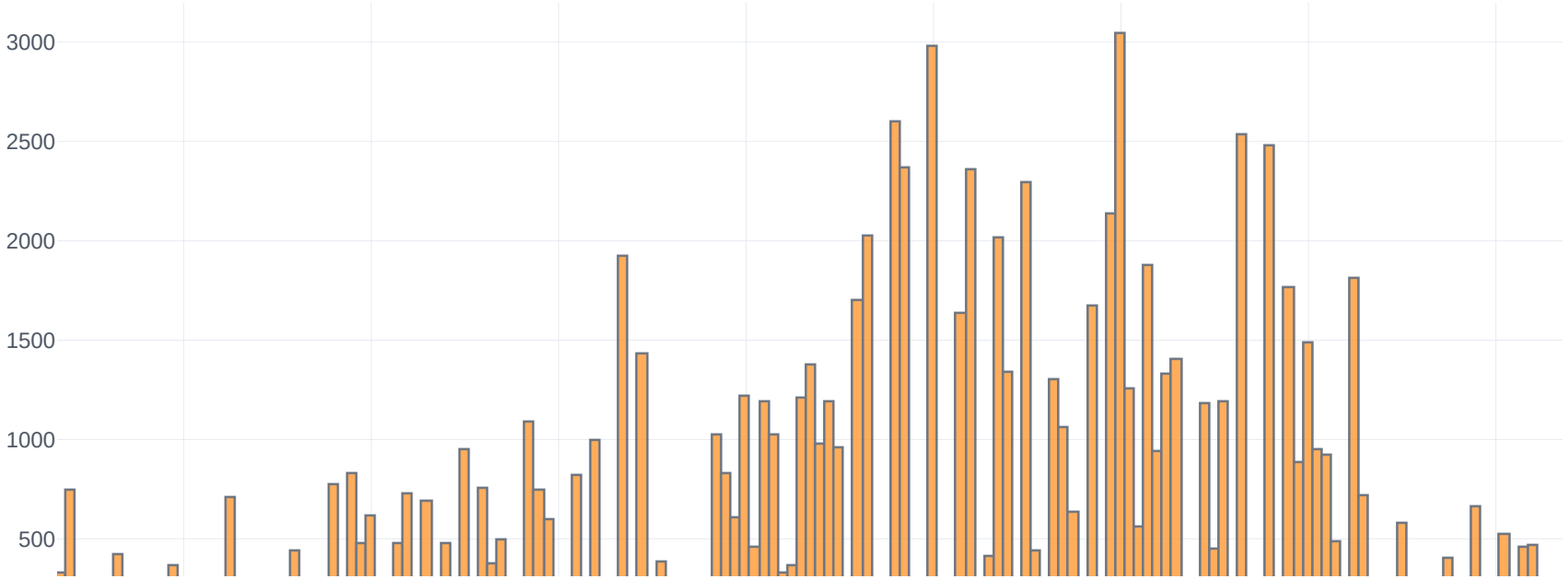

squad_size_local's histogram
14k
12k
10k
8k
6k
4k
2k


average_age_local's histogram
4500
4000
3500
3000
2500
2000
1500
1000
500

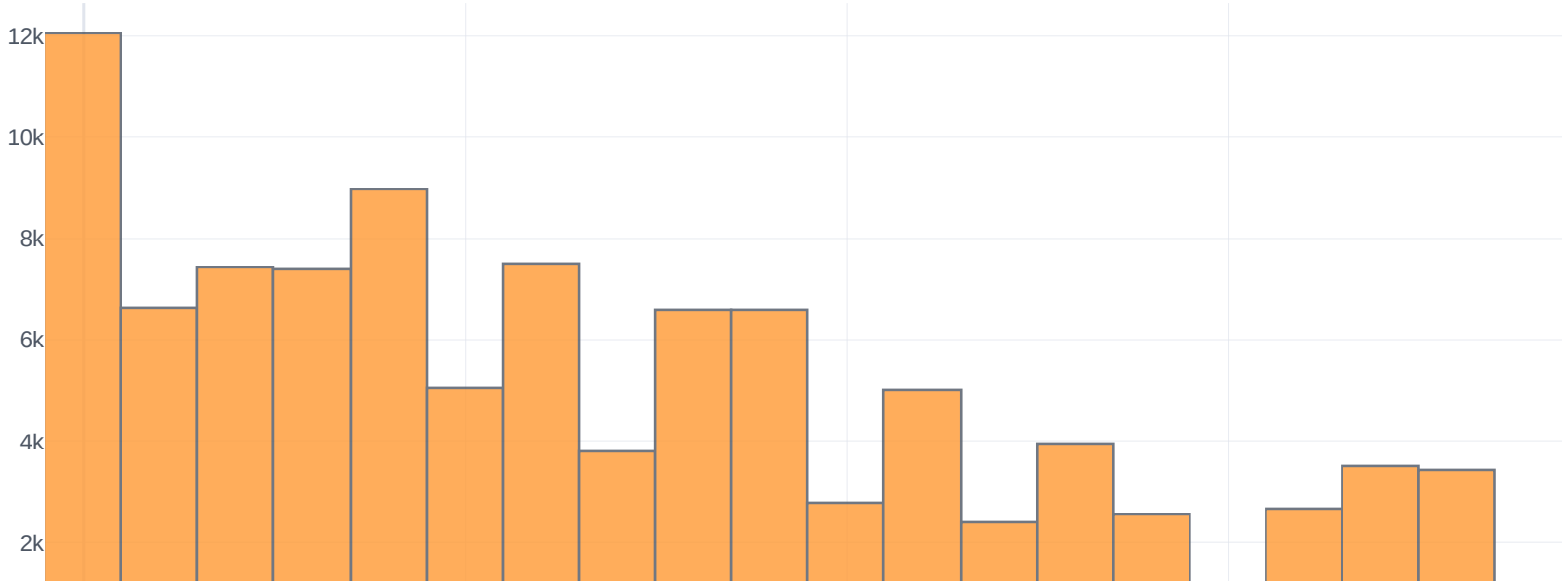
foreigners_number_local's histogram



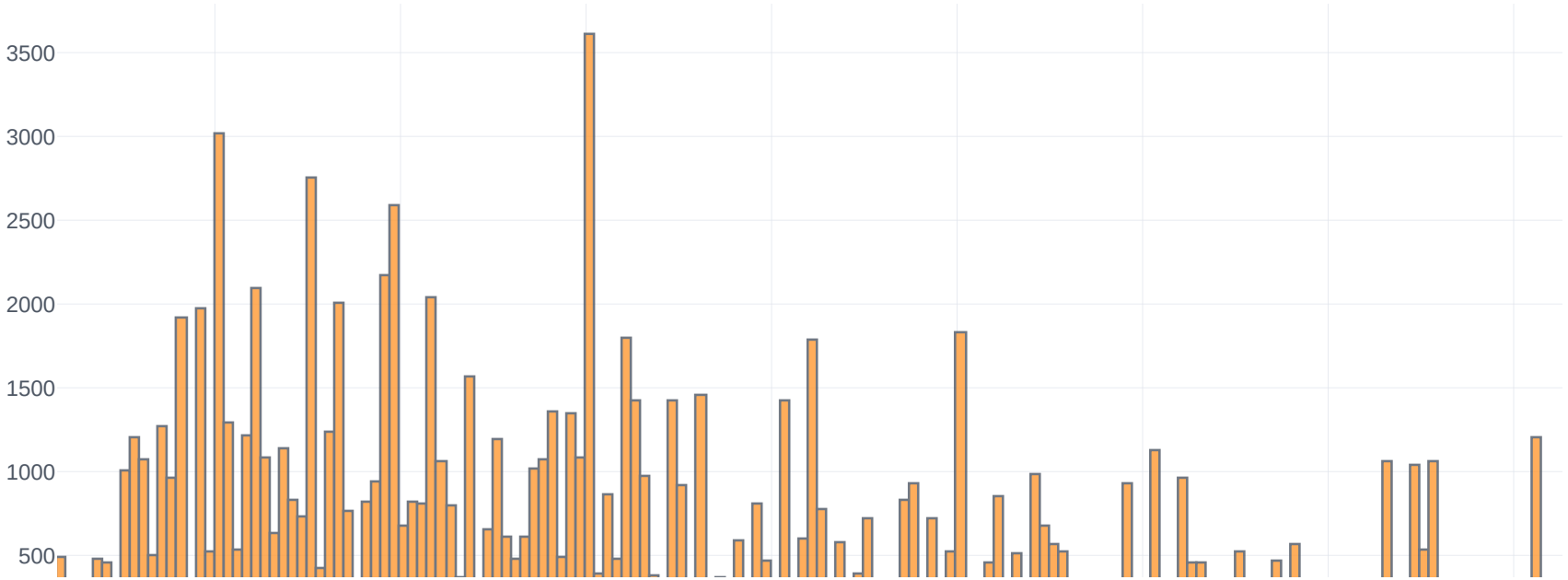
foreigners_percentage_local's histogram

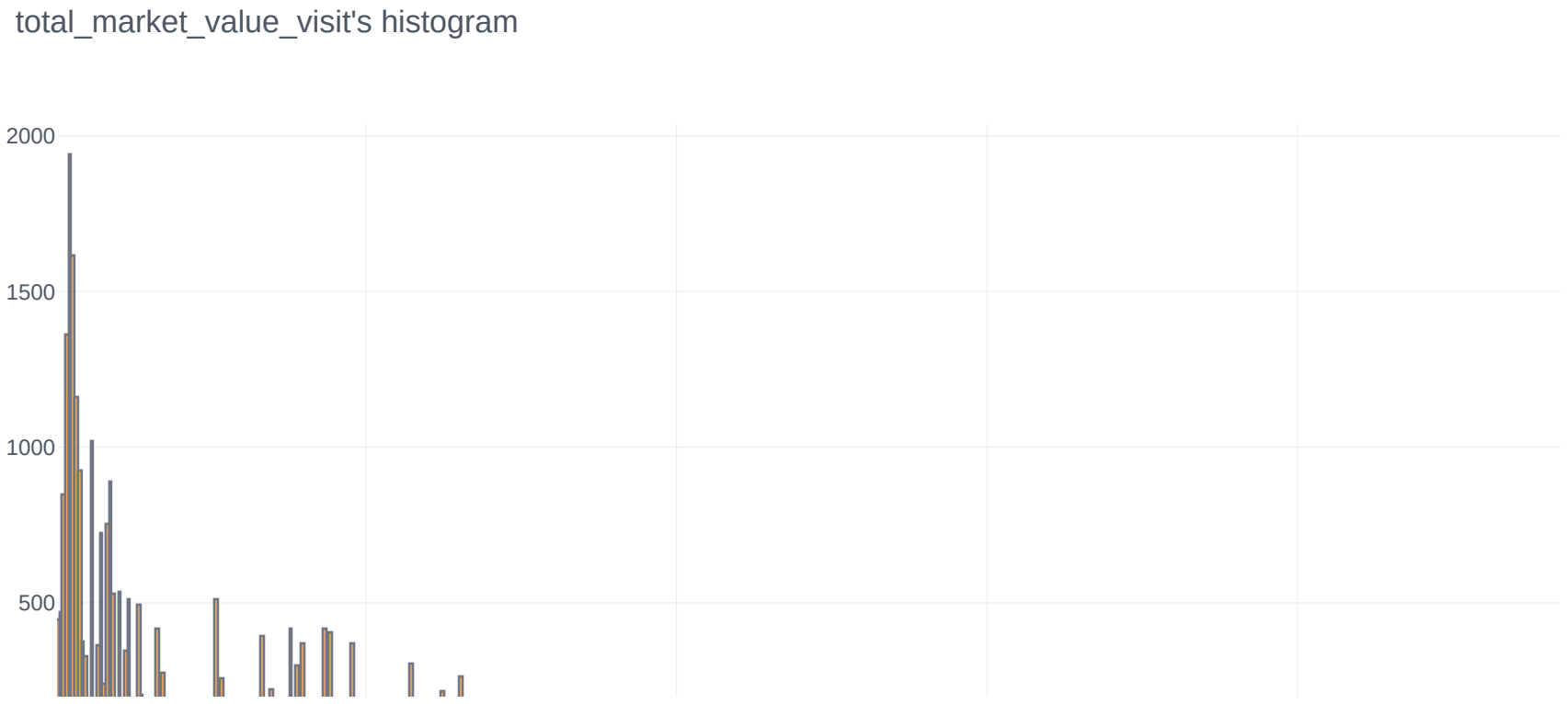


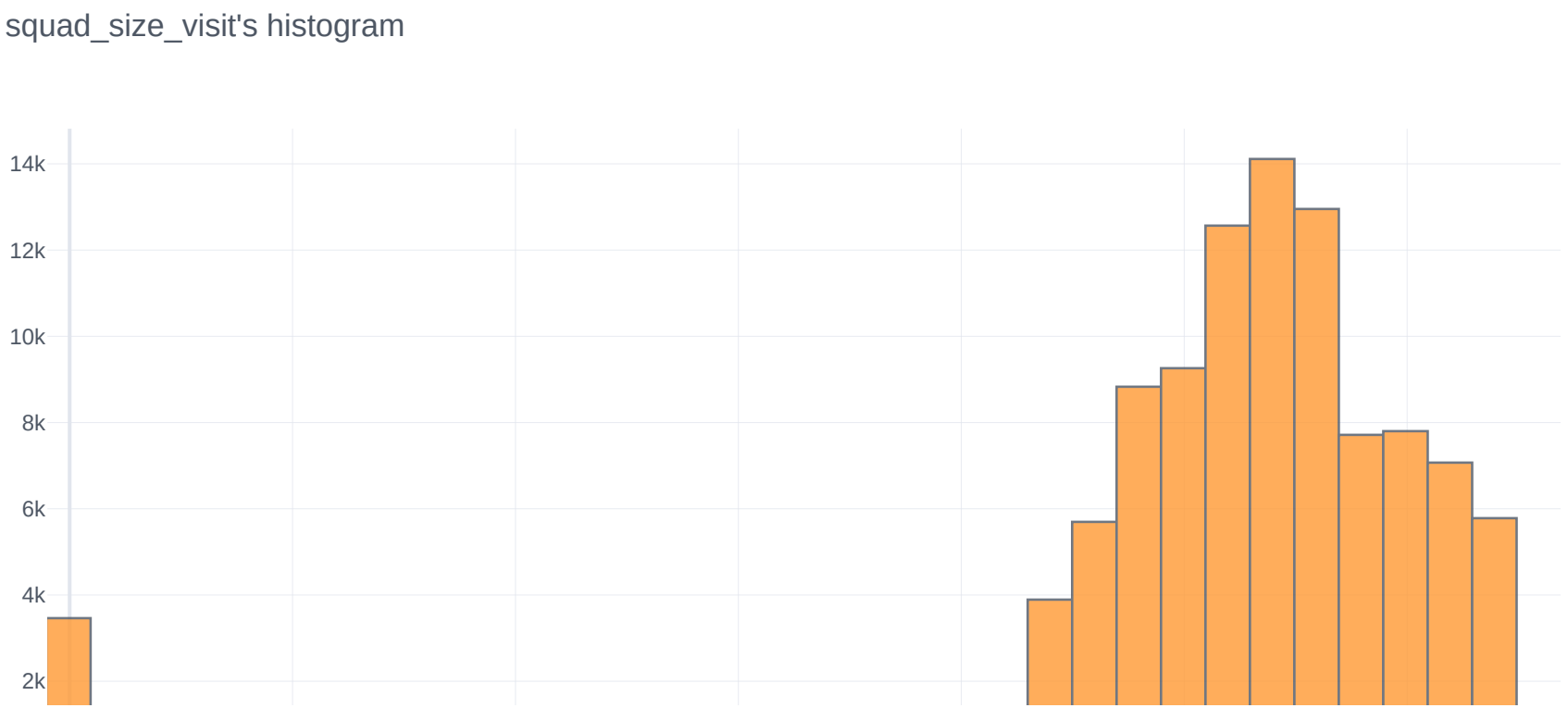
national_team_players_local's histogram

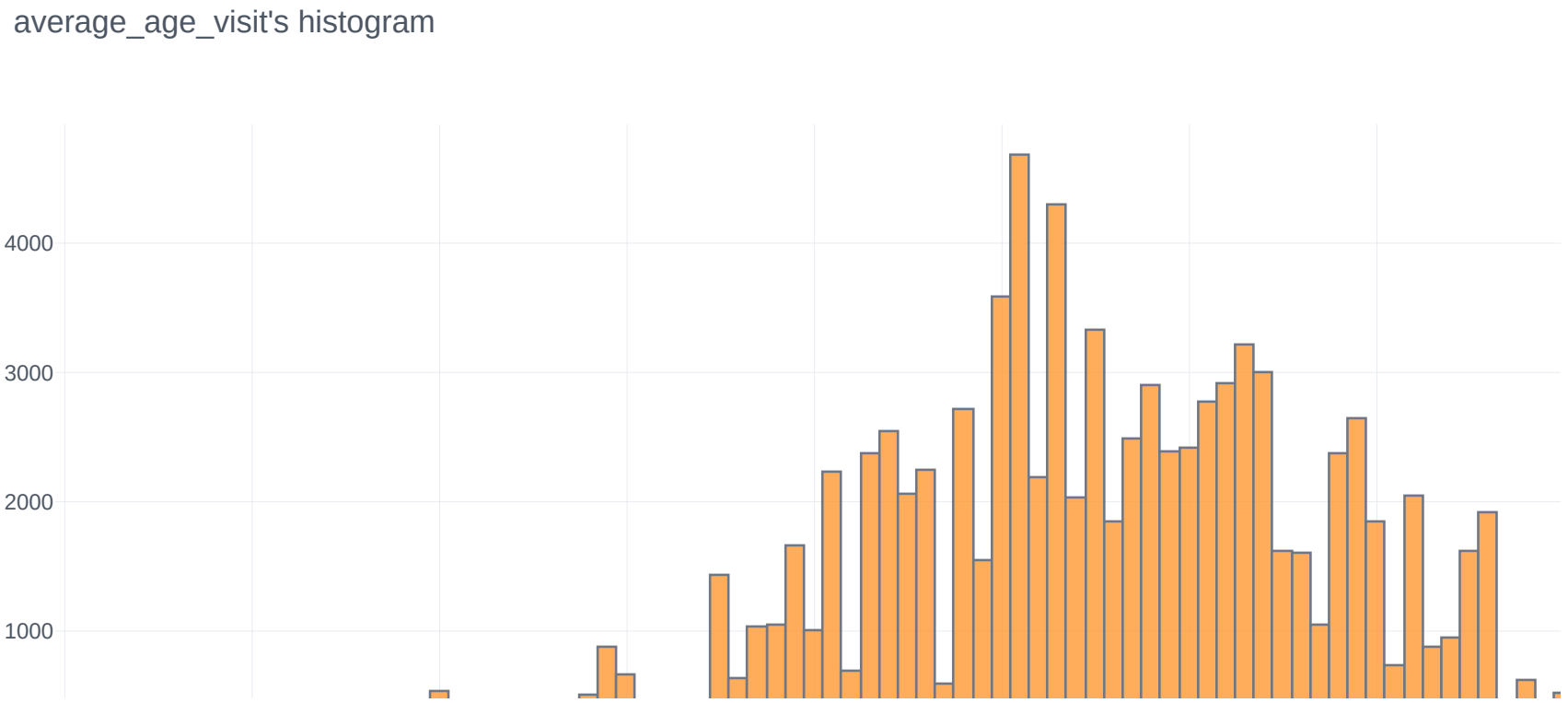


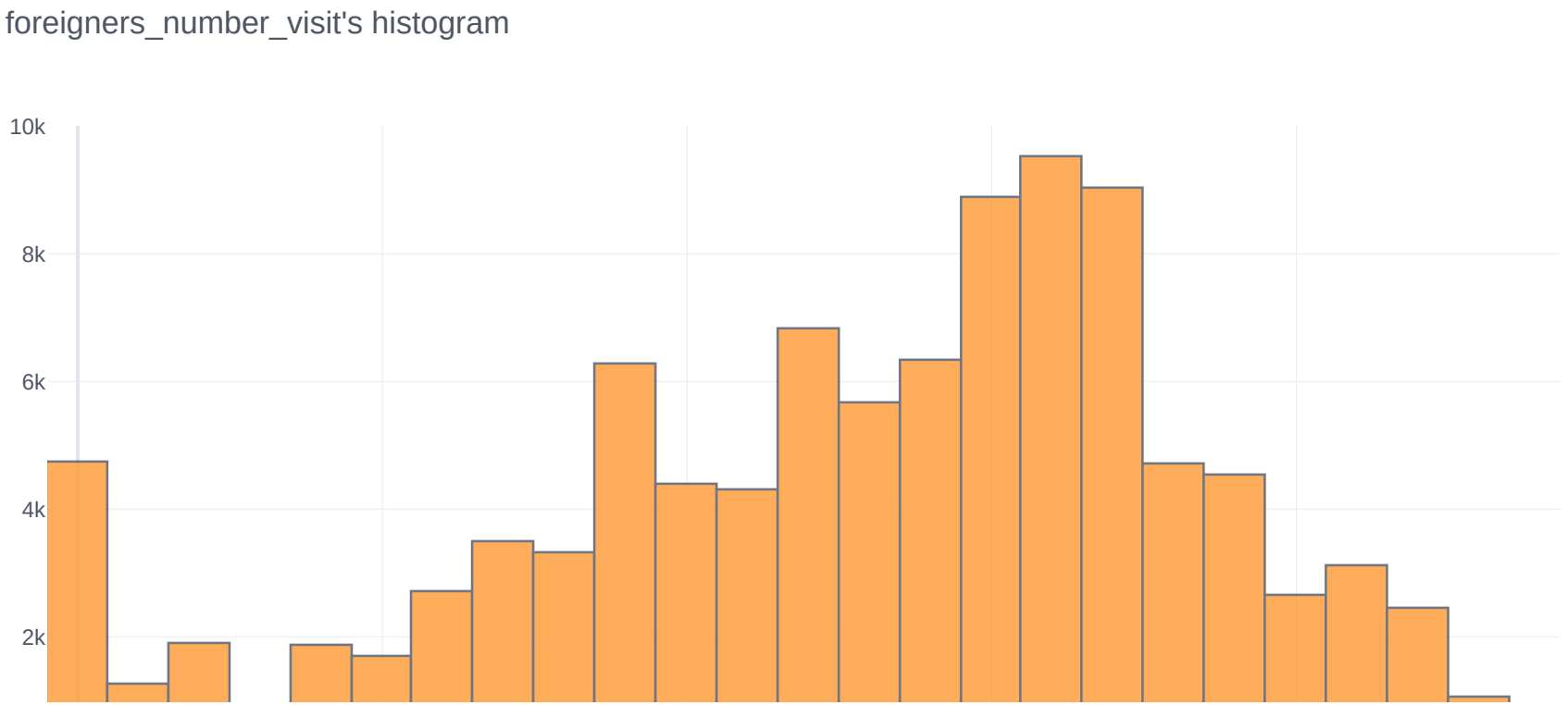
stadium_seats_local's histogram



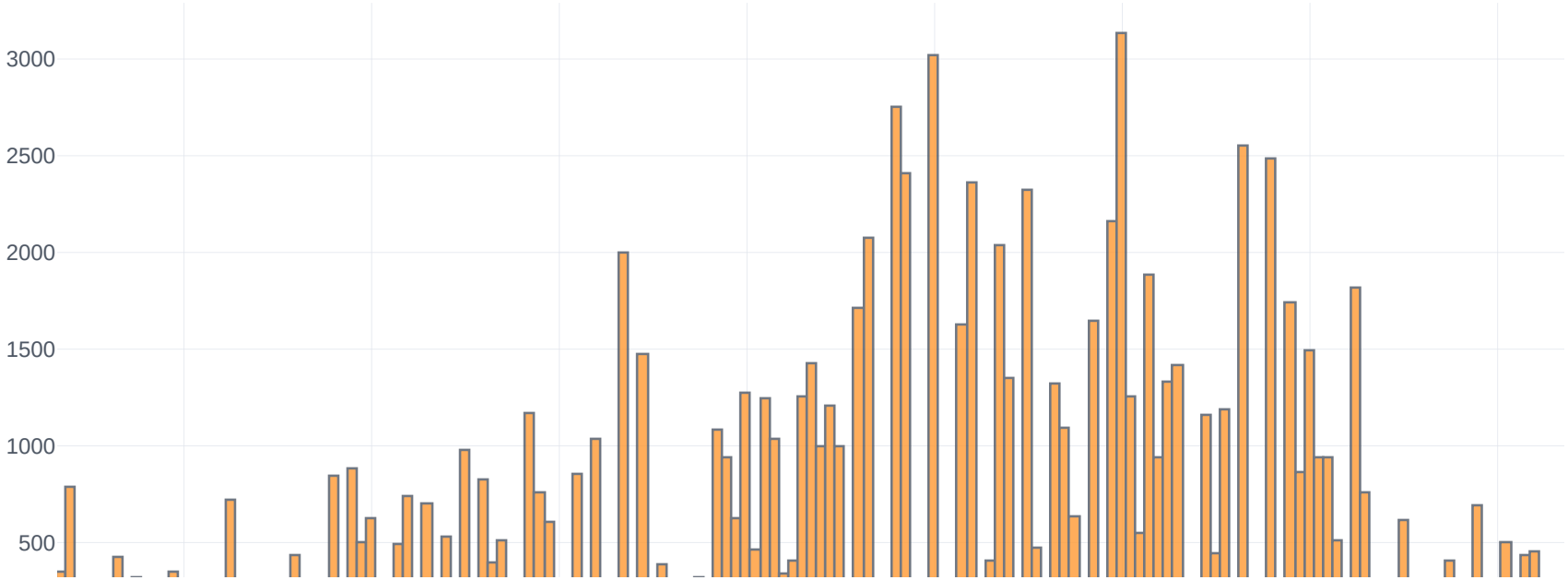

total_market_value_visit's histogram
2000
1500
1000
500


squad_size_visit's histogram
14k
12k
10k
8k
6k
4k
2k

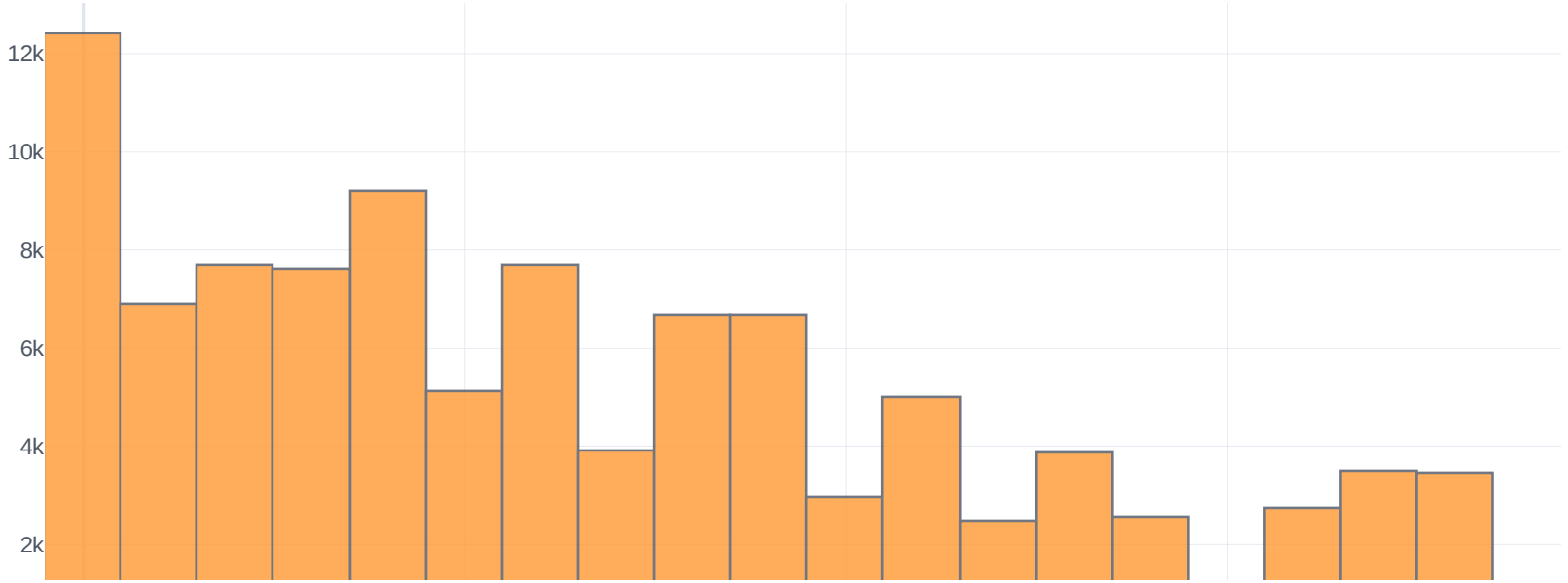

average_age_visit's histogram
4000
3000
2000
1000

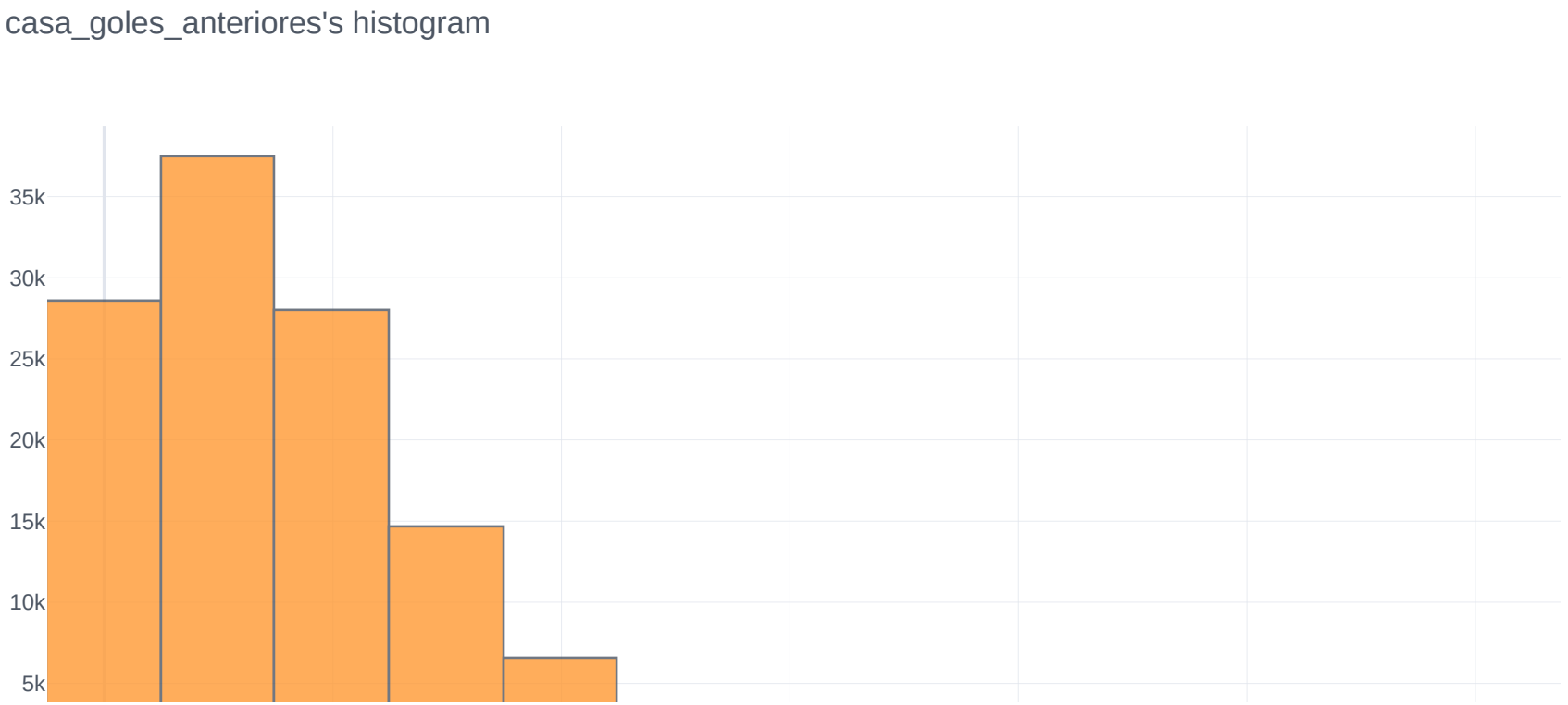

foreigners_number_visit's histogram
10k
8k
6k
4k
2k

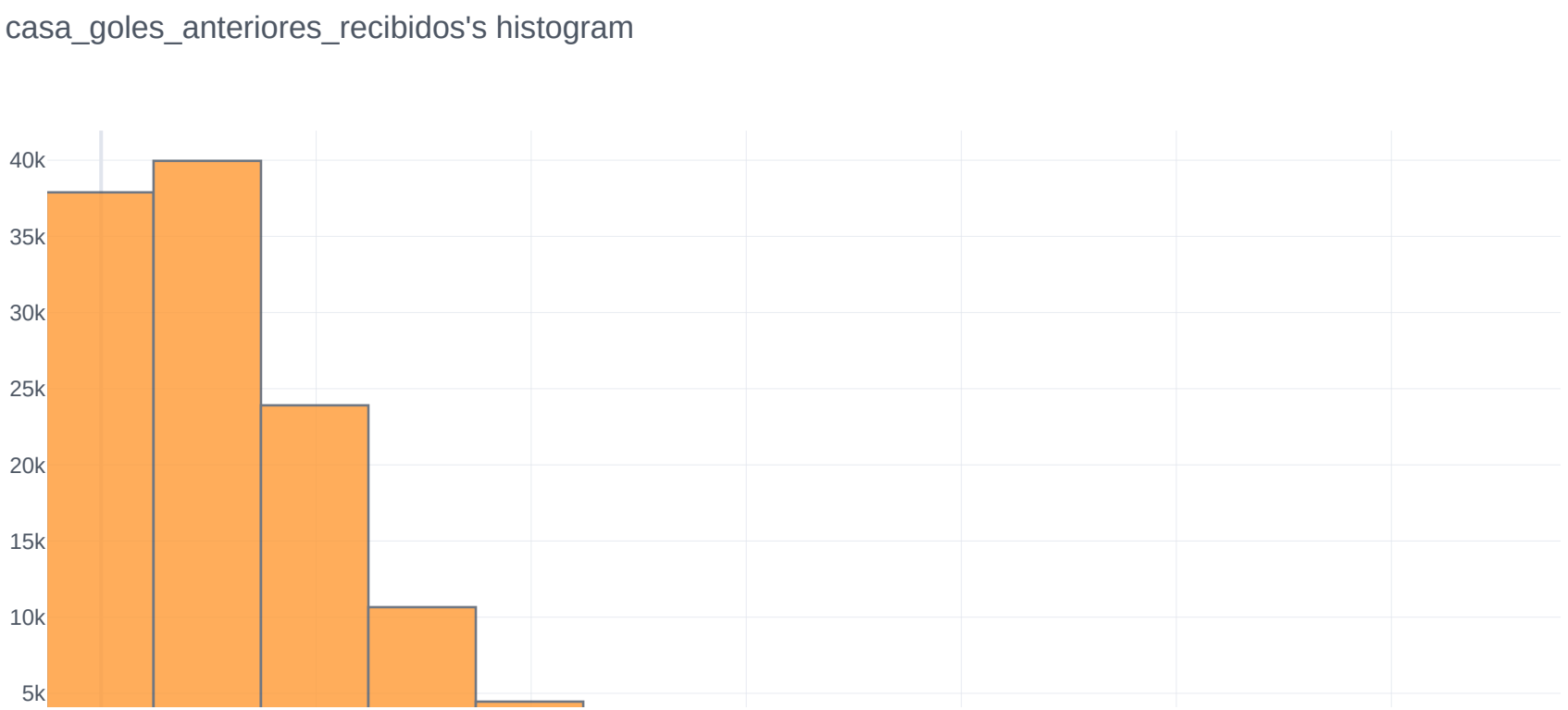
foreigners_percentage_visit's histogram

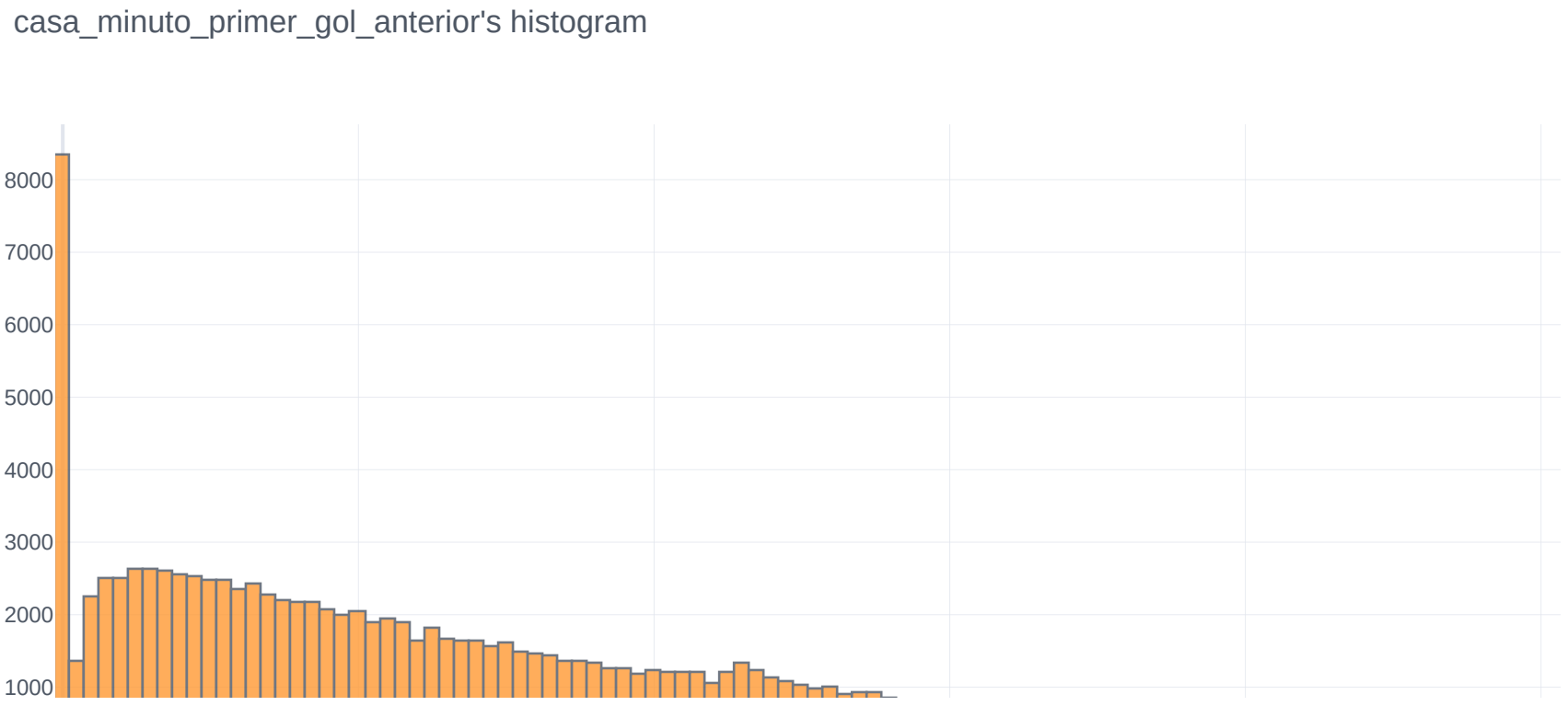


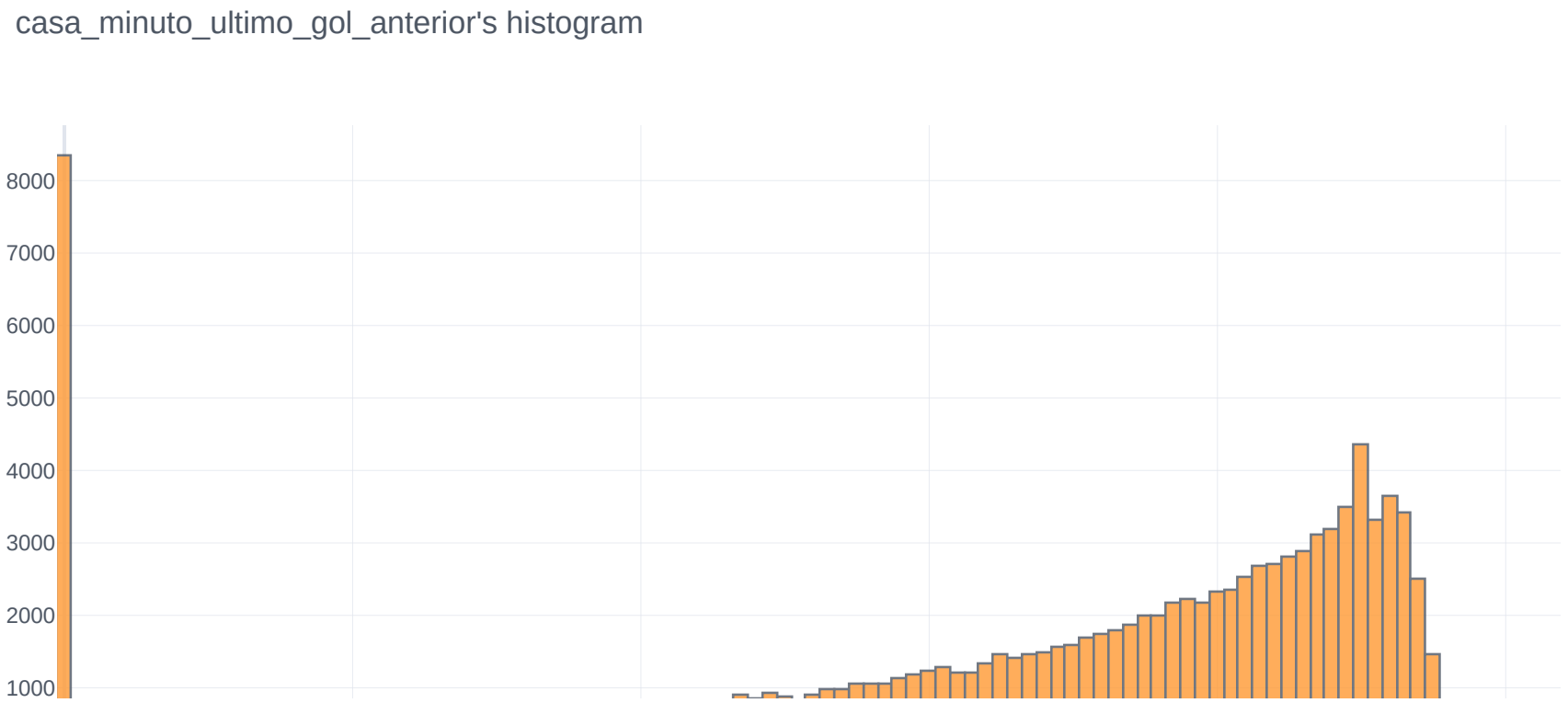
national_team_players_visit's histogram



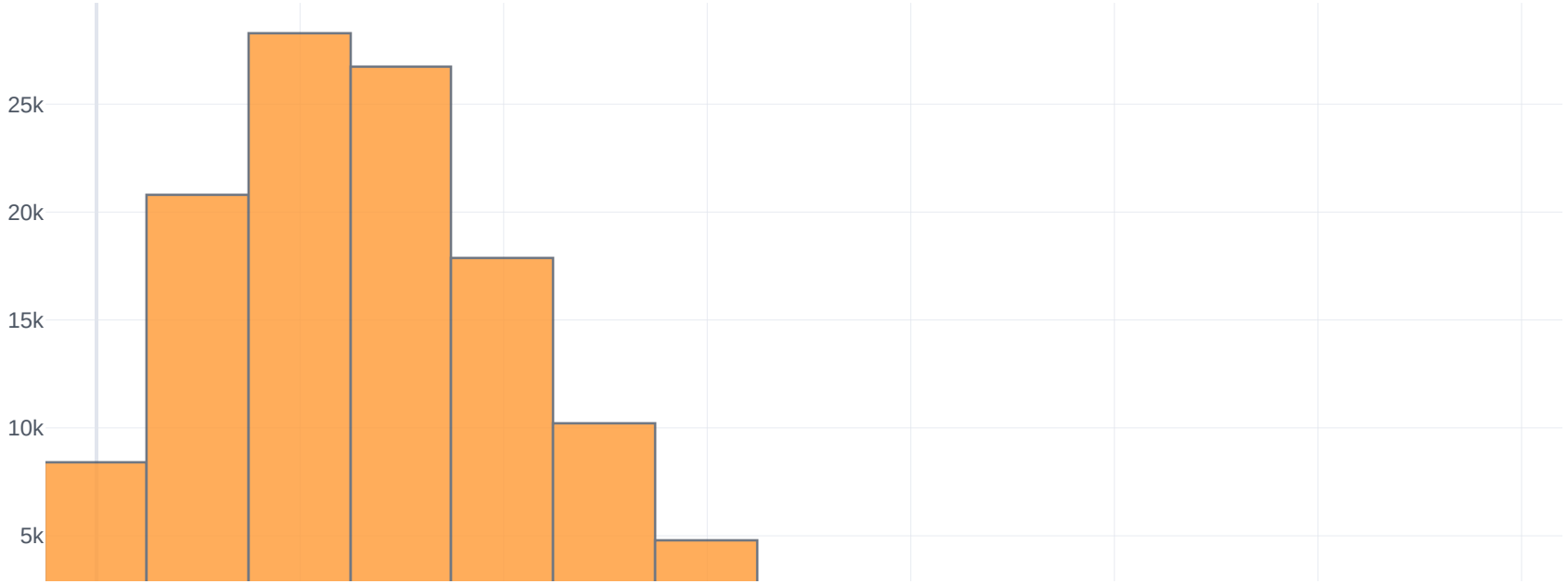

casa_goles_anteriores's histogram
35k
30k
25k
20k
15k
10k
5k

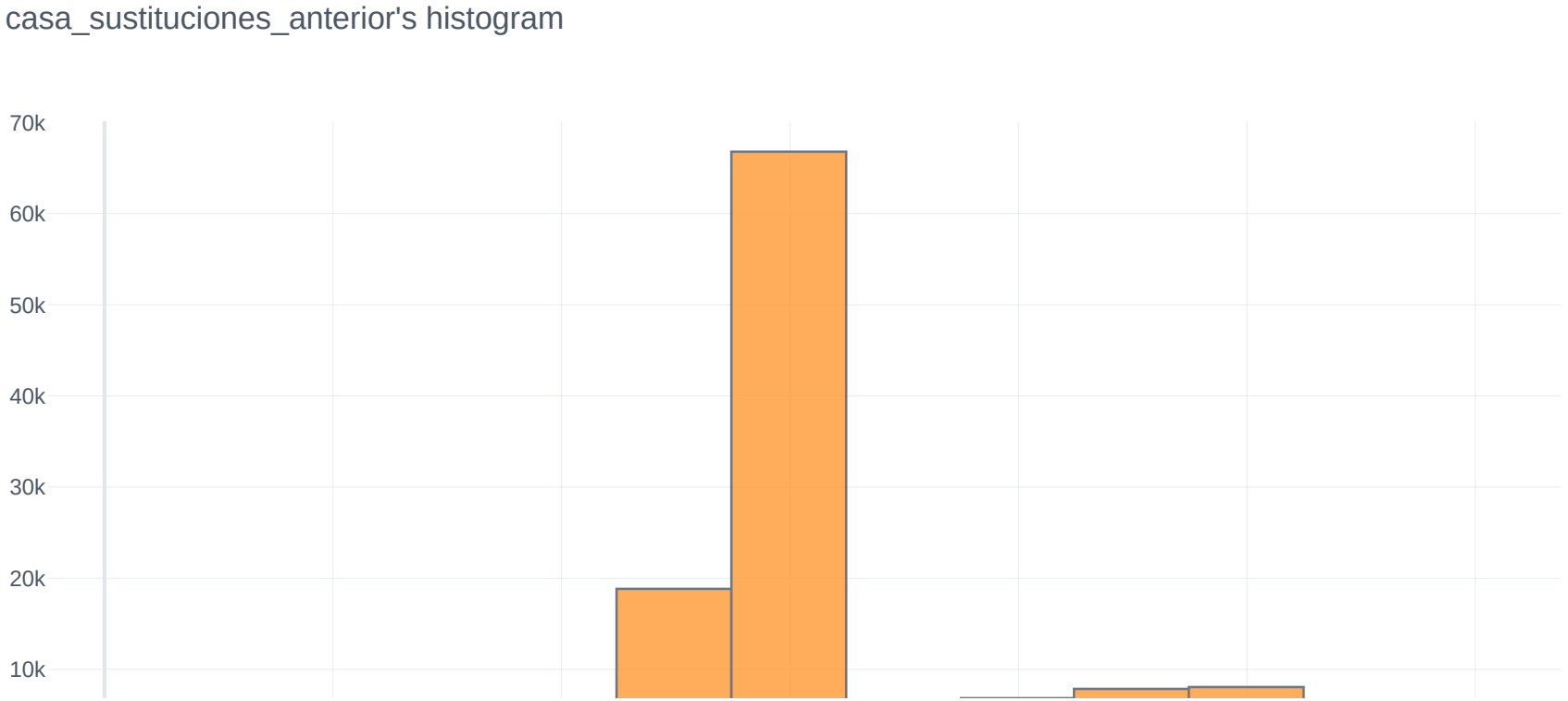

casa_goles_anteriores_recibidos's histogram
40k
35k
30k
25k
20k
15k
10k
5k


casa_minuto_primer_gol_anterior's histogram
8000
7000
6000
5000
4000
3000
2000
1000

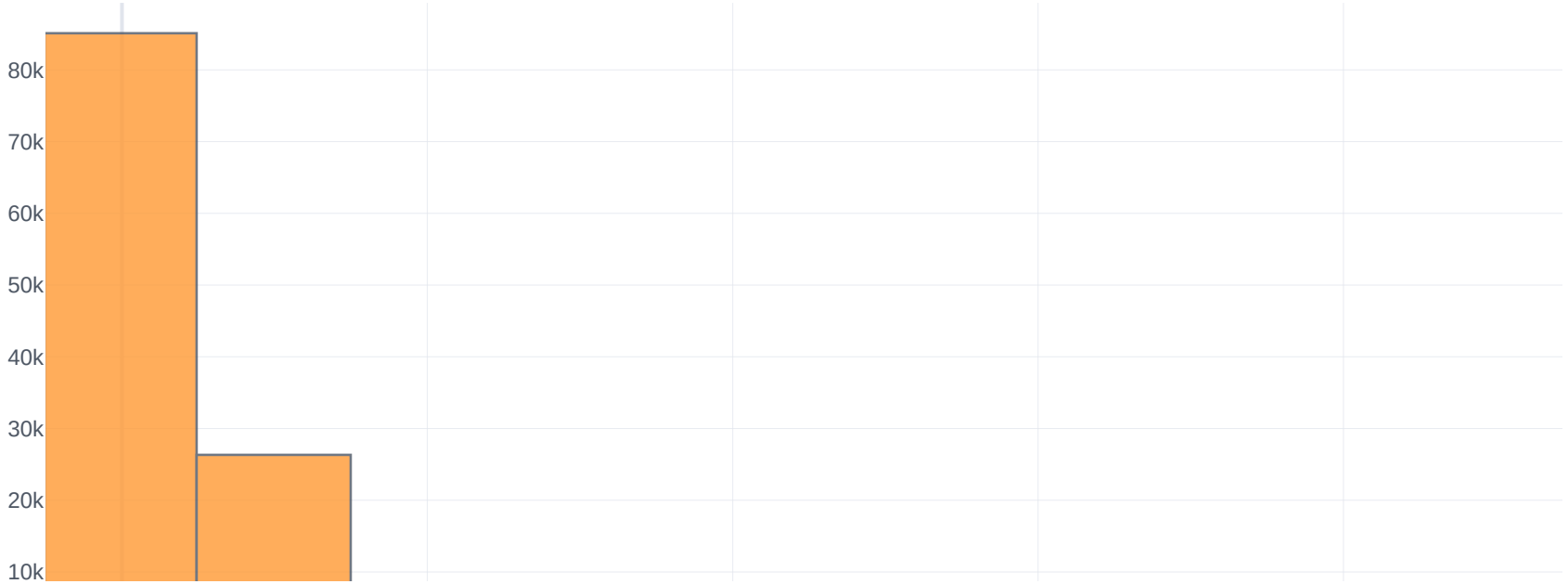

casa_minuto_ultimo_gol_anterior's histogram
8000
7000
6000
5000
4000
3000
2000
1000

casa_number_goals_anterior's histogram

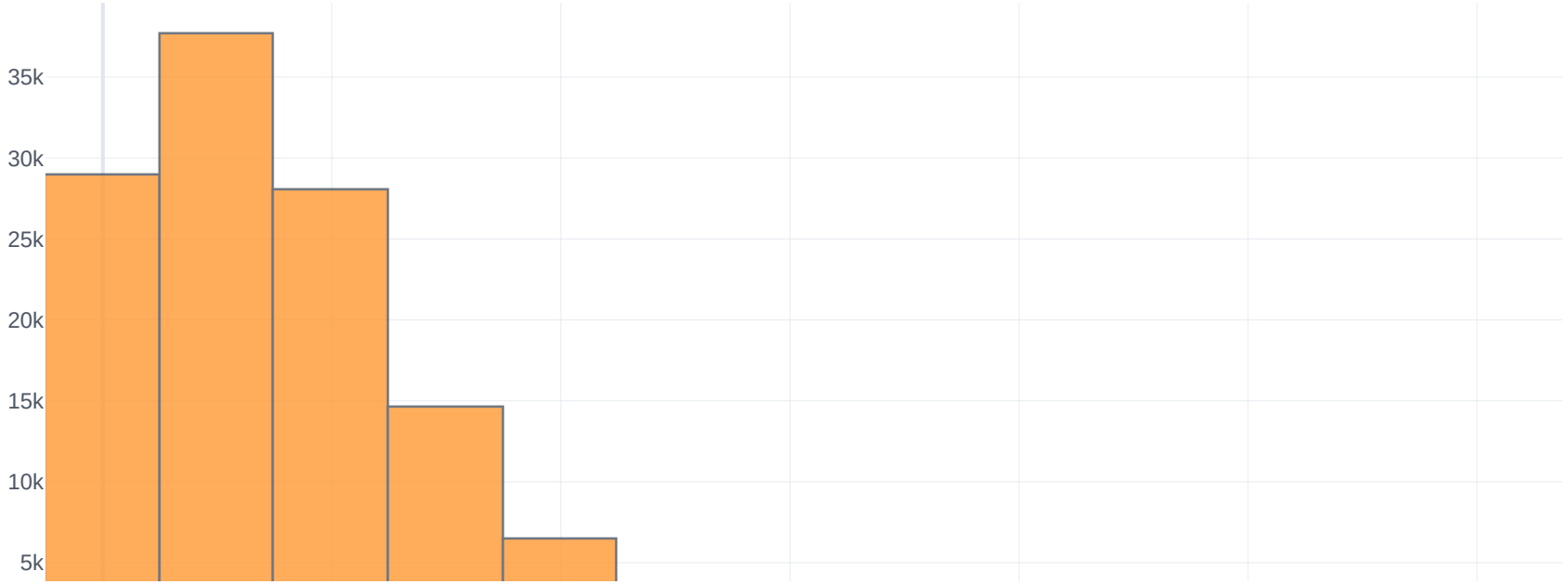



casa_sustituciones_anterior's histogram
70k
60k
50k
40k
30k
20k
10k

casa_goles_tiempo_añadido_anterior's histogram

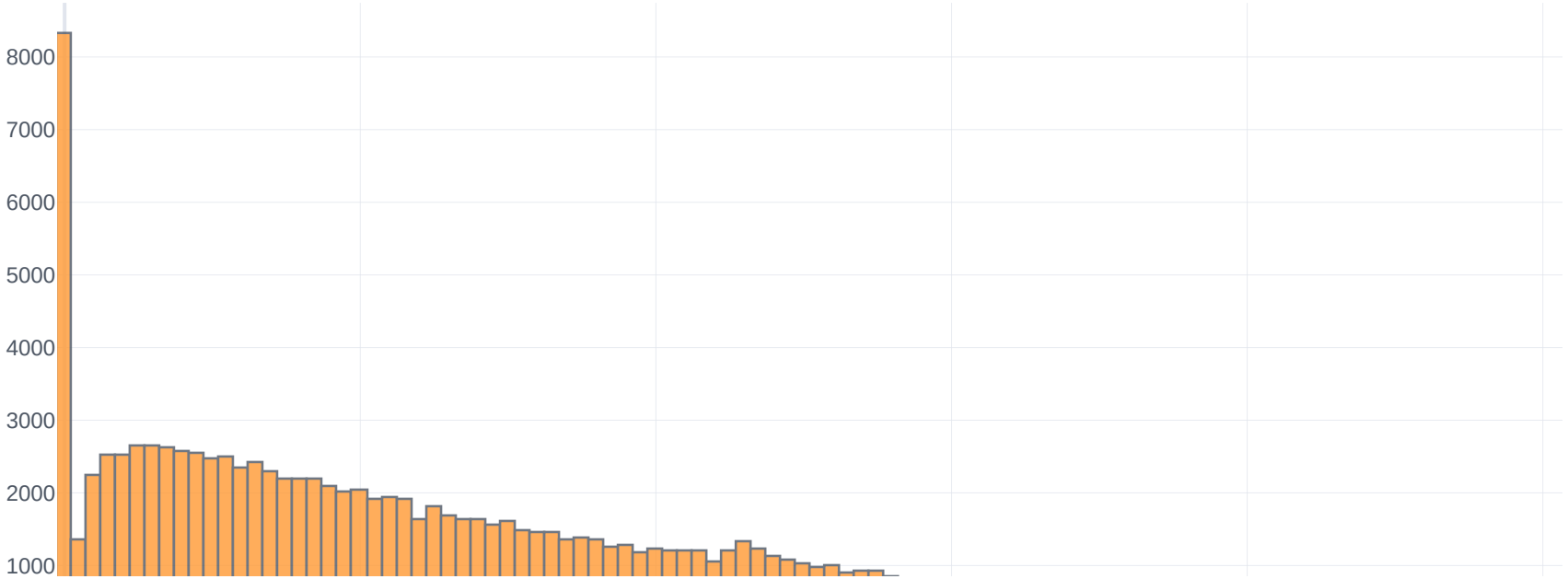


visita_goles_anteriores_recibidos's histogram

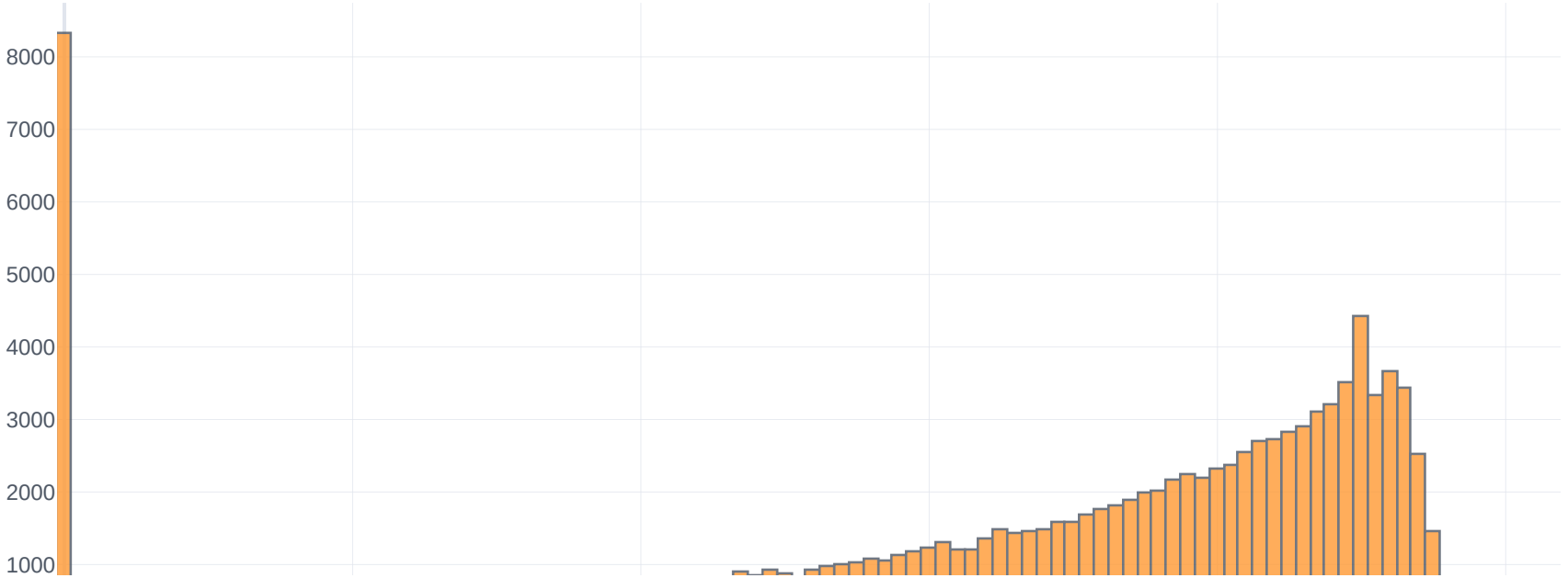


visita_goles_anteriores's histogram

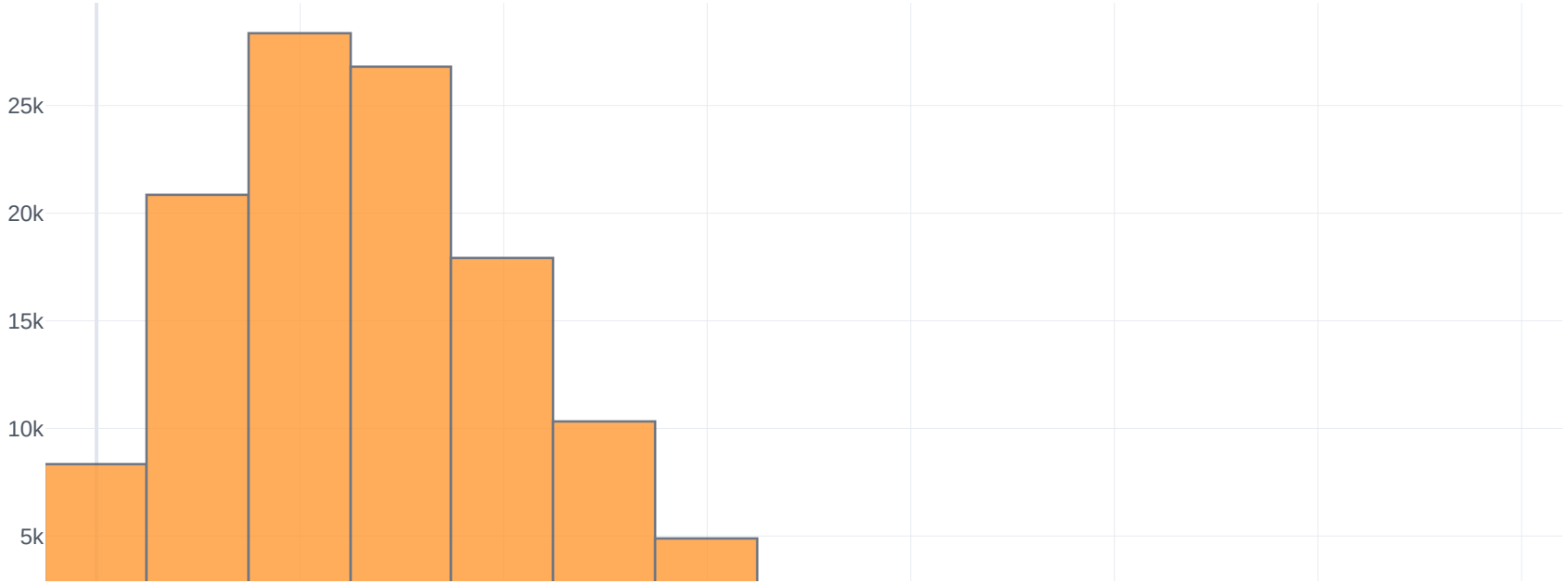
visita_minuto_primer_gol_anterior's histogram



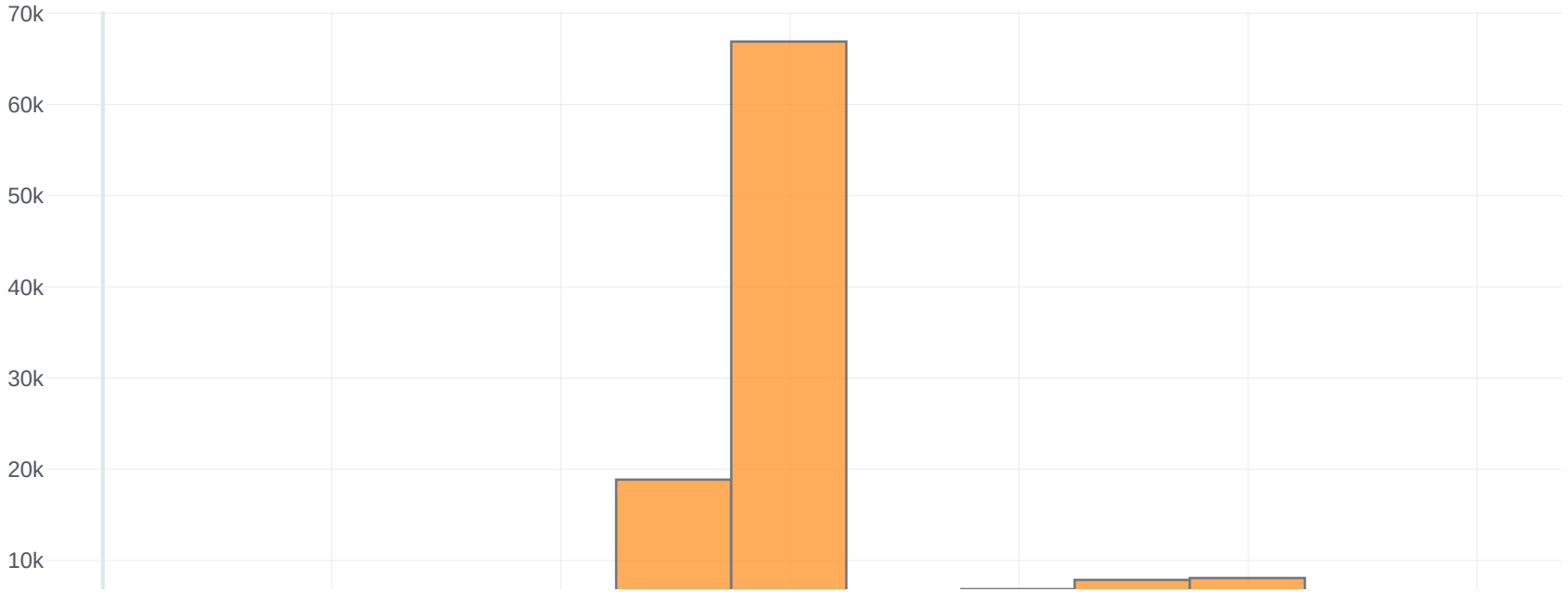
visita_minuto_ultimo_gol_anterior's histogram



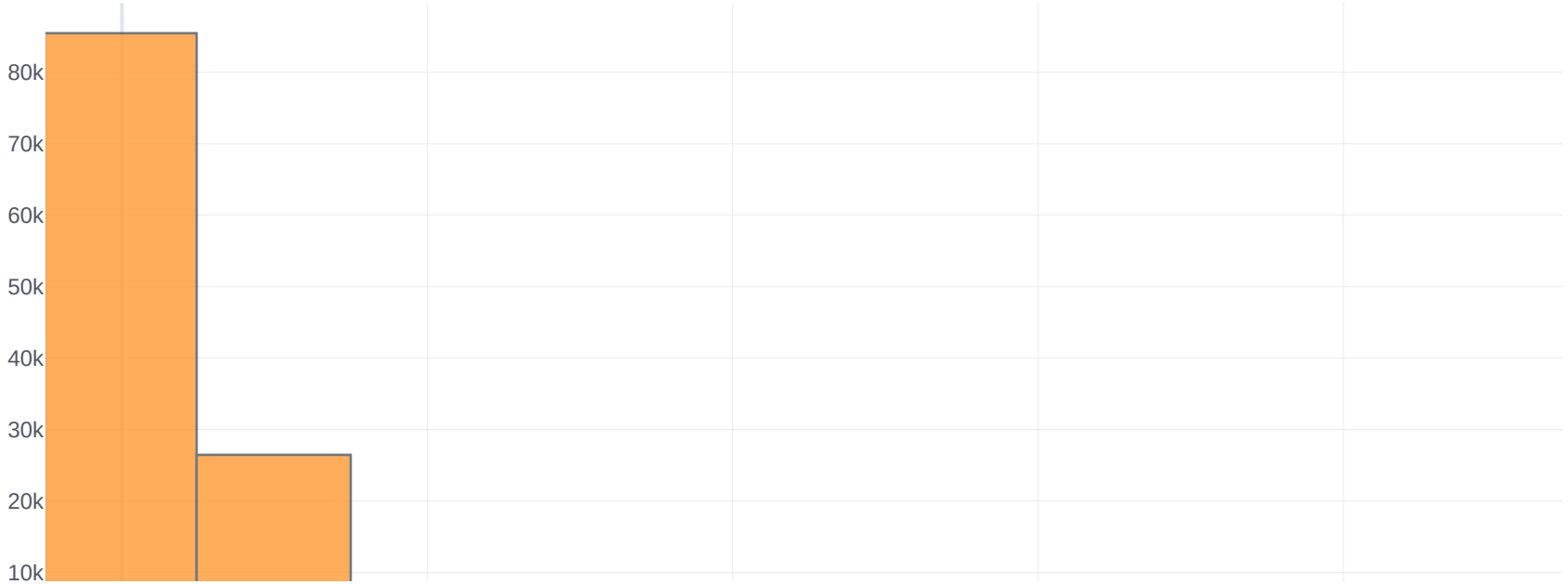
visita_number_goals_anterior's histogram

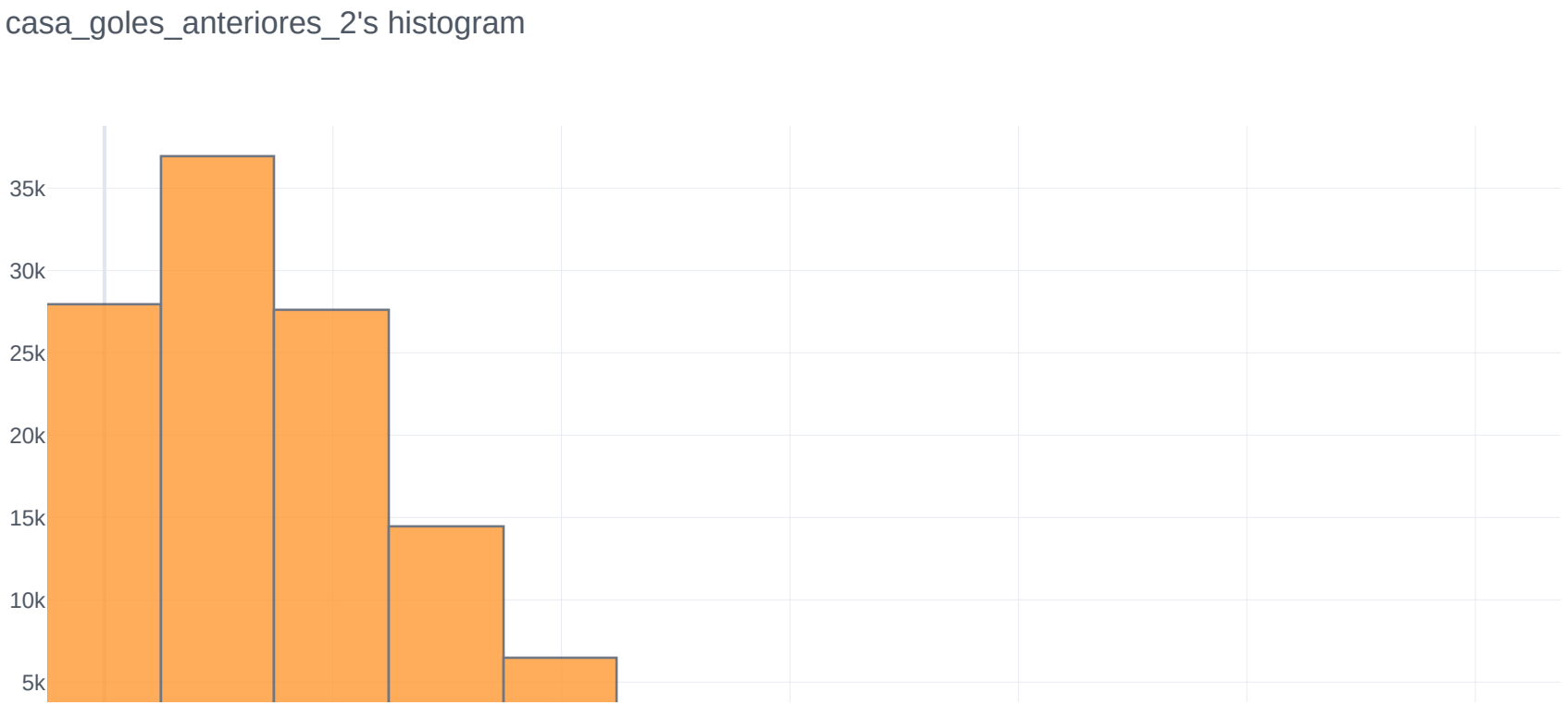


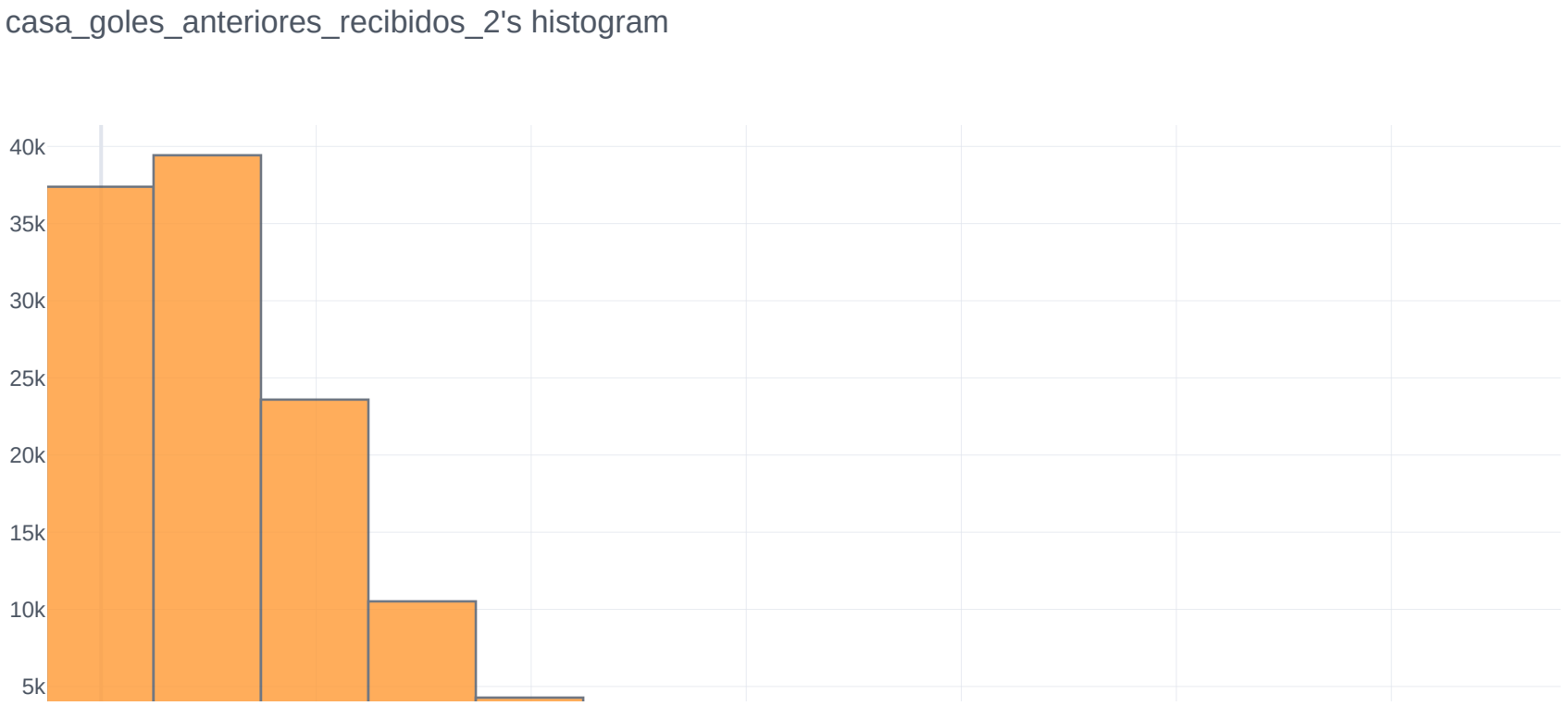
visita_sustituciones_anterior's histogram

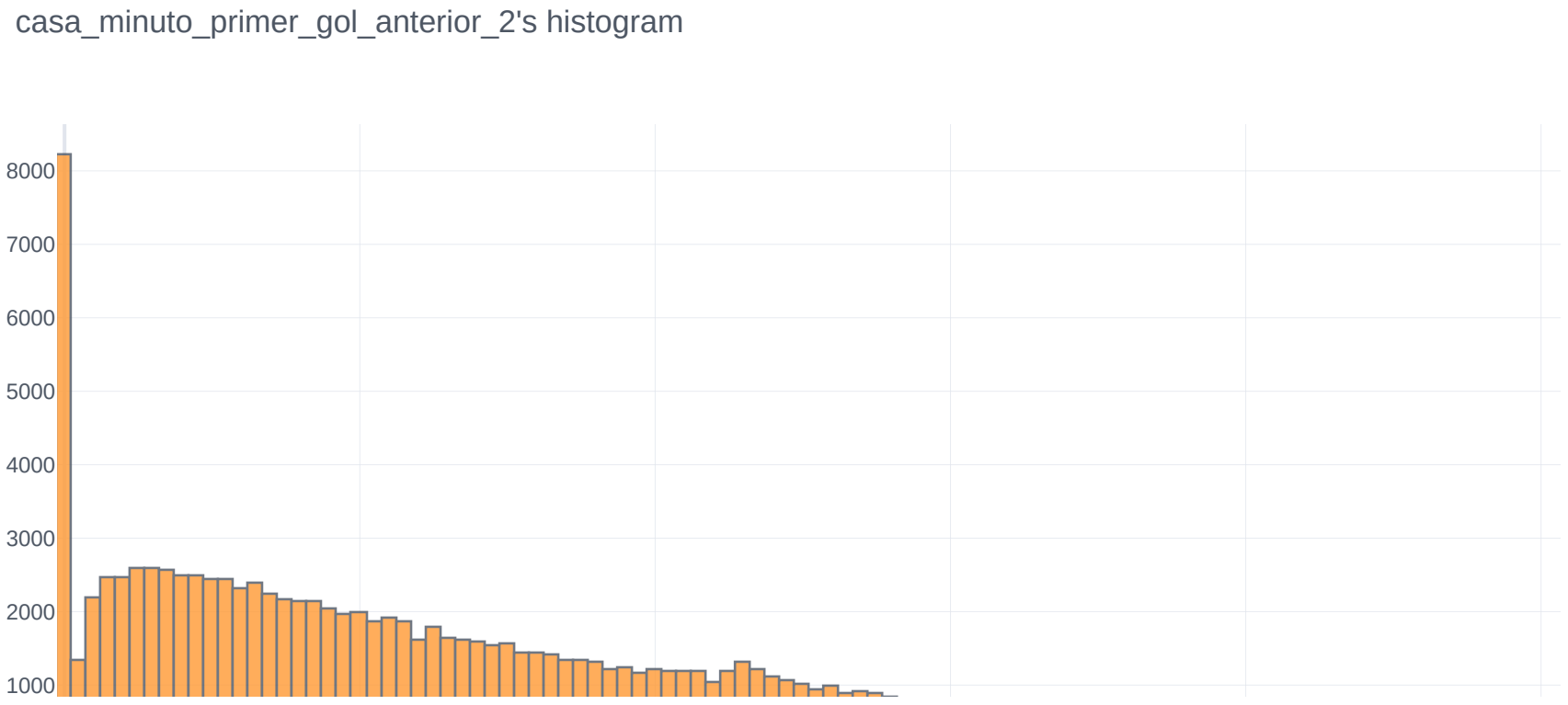


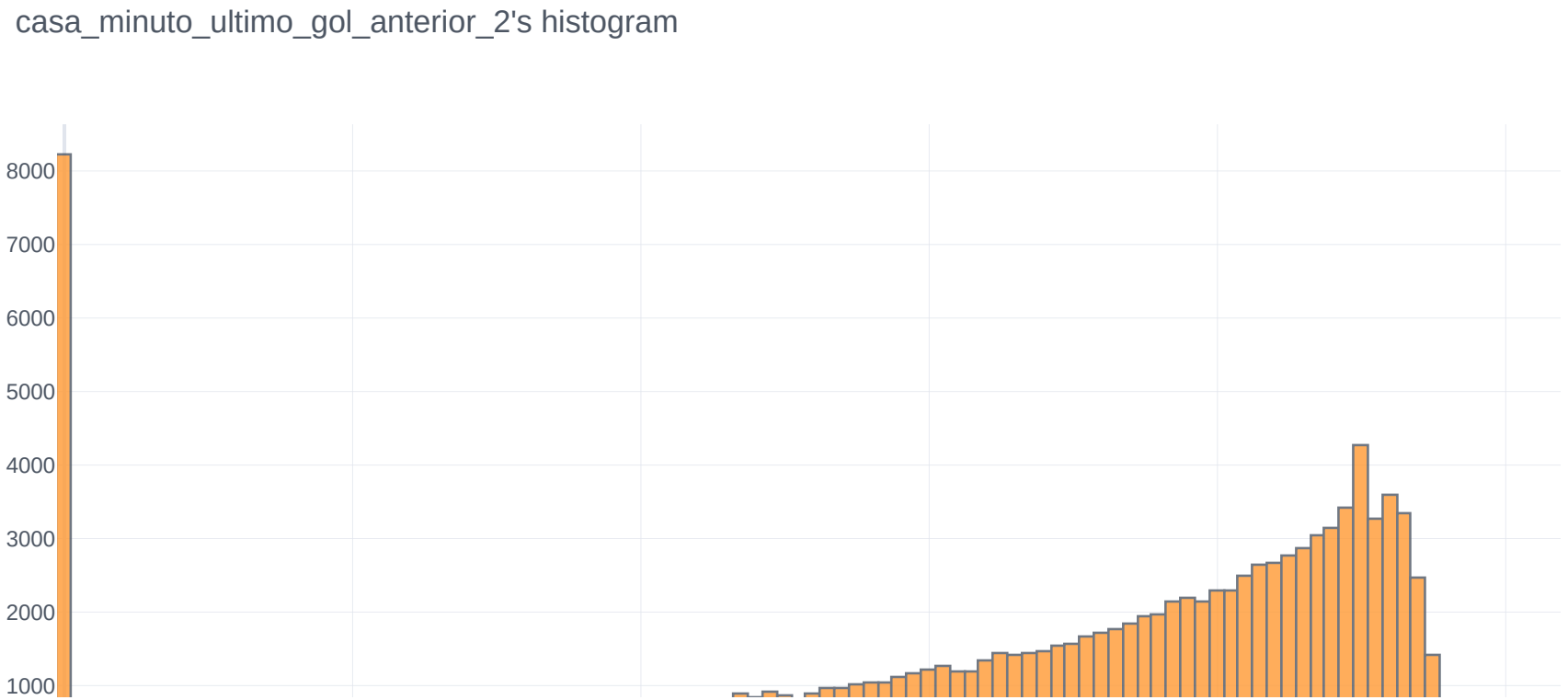
visita_goles_tiempo_añadido_anterior's histogram

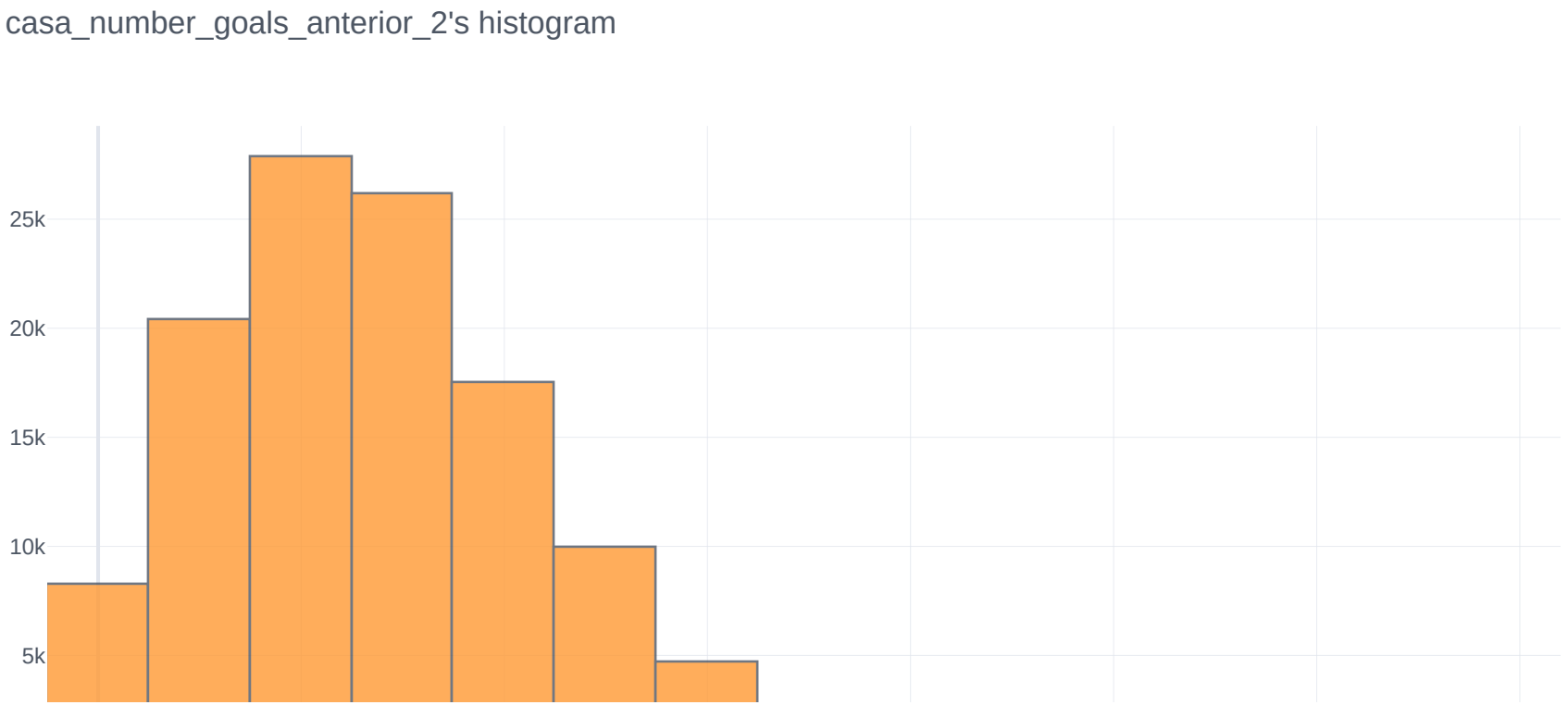


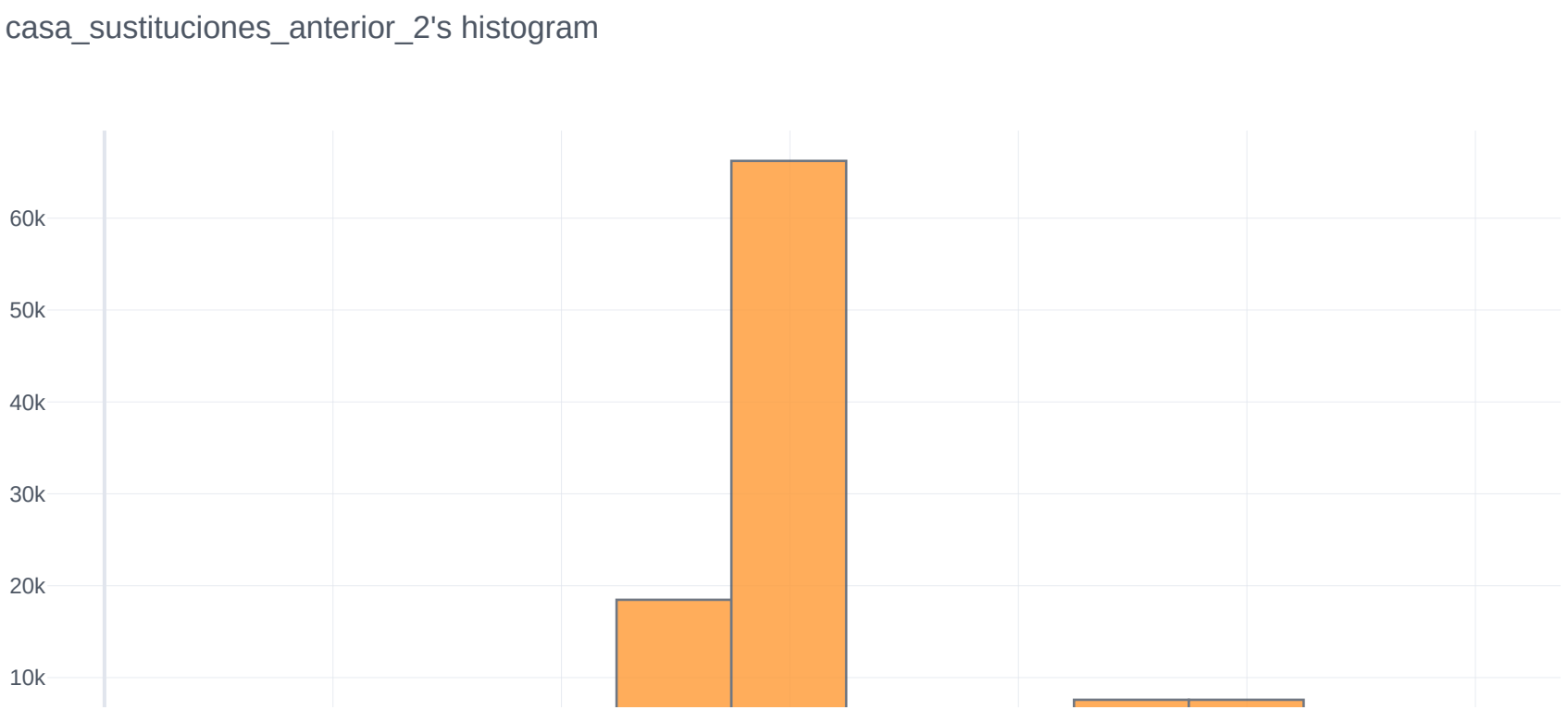

casa_goles_anteriores_2's histogram
35k
30k
25k
20k
15k
10k
5k


casa_goles_anteriores_recibidos_2's histogram
40k
35k
30k
25k
20k
15k
10k
5k


casa_minuto_primer_gol_anterior_2's histogram
8000
7000
6000
5000
4000
3000
2000
1000

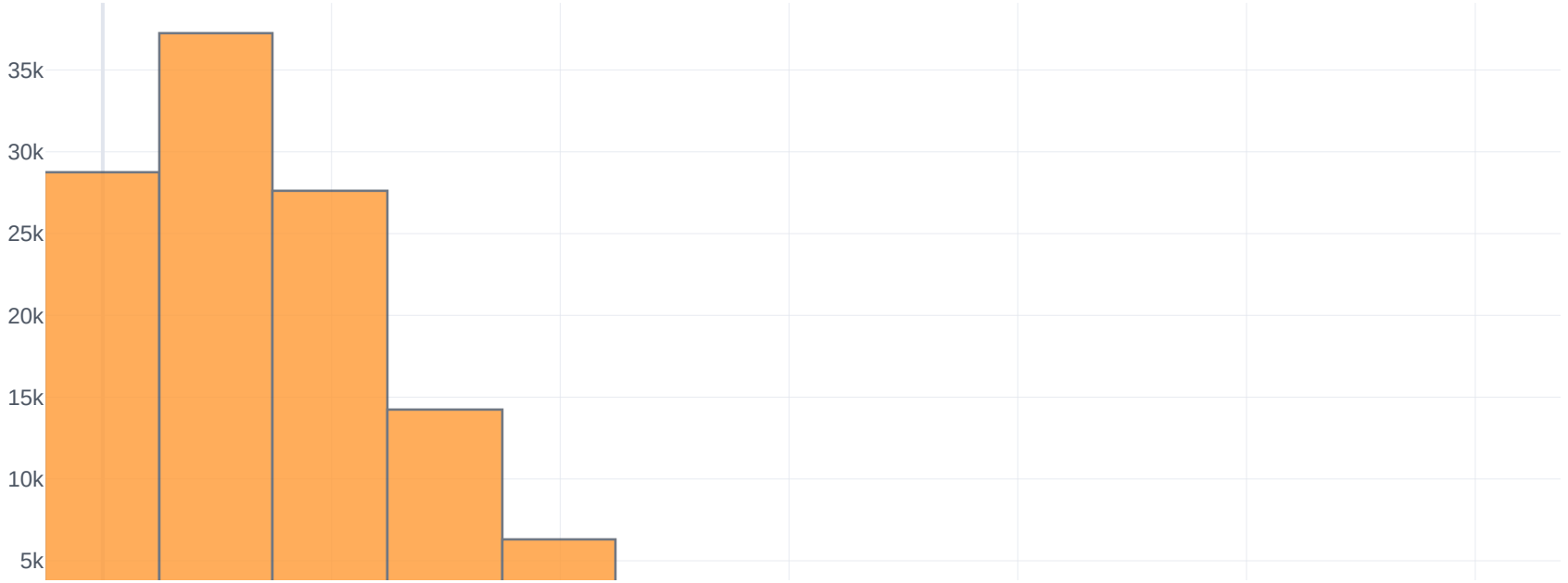

casa_minuto_ultimo_gol_anterior_2's histogram
8000
7000
6000
5000
4000
3000
2000
1000


casa_number_goals_anterior_2's histogram
25k
20k
15k
10k
5k


casa_sustituciones_anterior_2's histogram
60k
50k
40k
30k
20k
10k

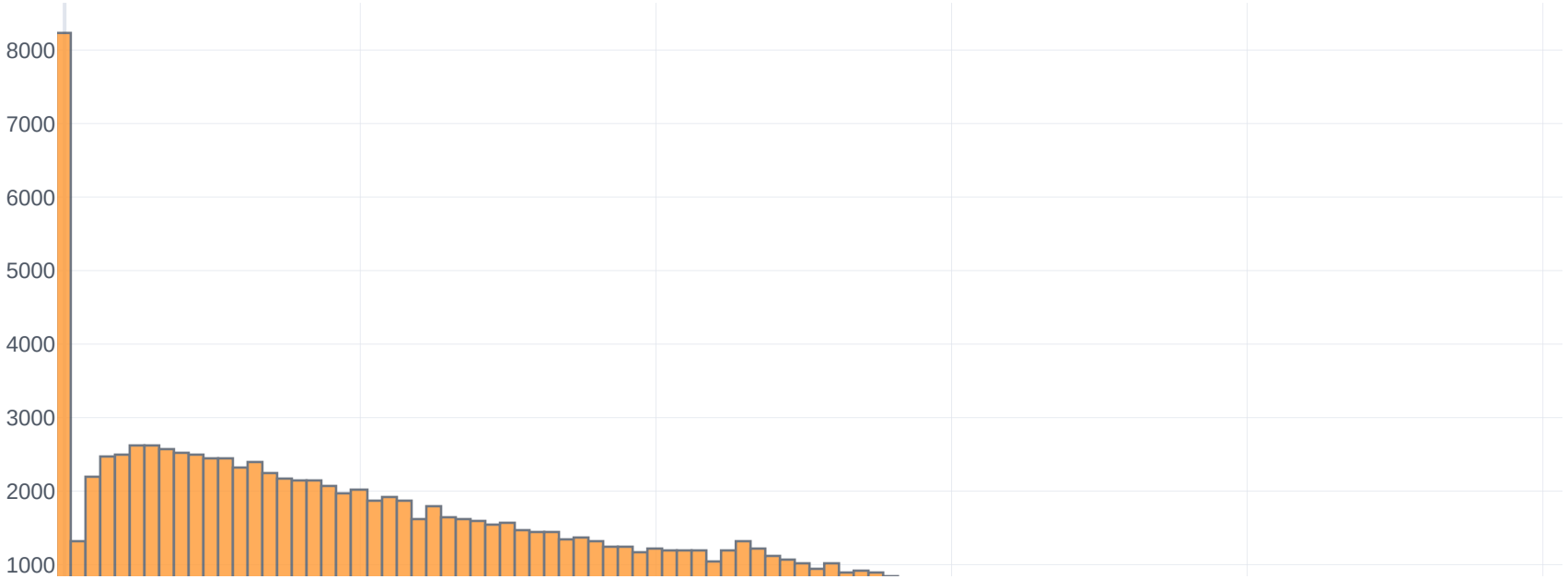
casa_goles_tiempo_añadido_anterior_2's histogram

visita_goles_anteriores_recibidos_2's histogram

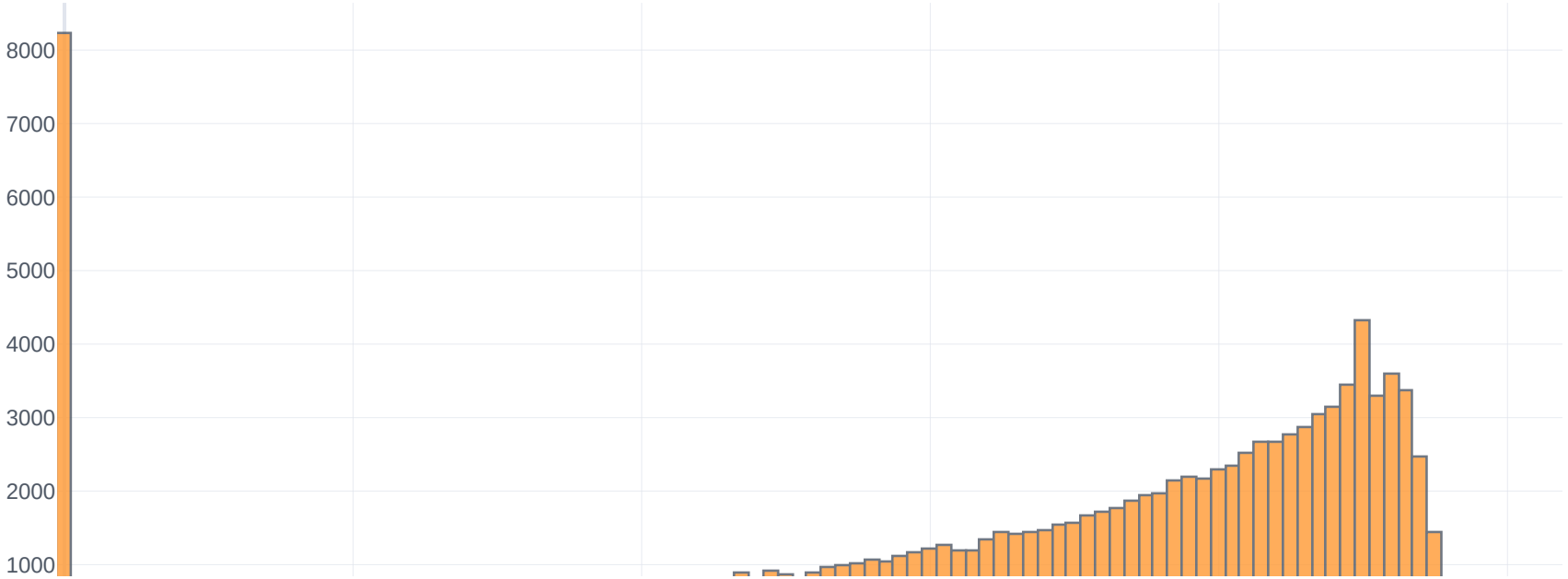


visita_goles_anteriores_2's histogram

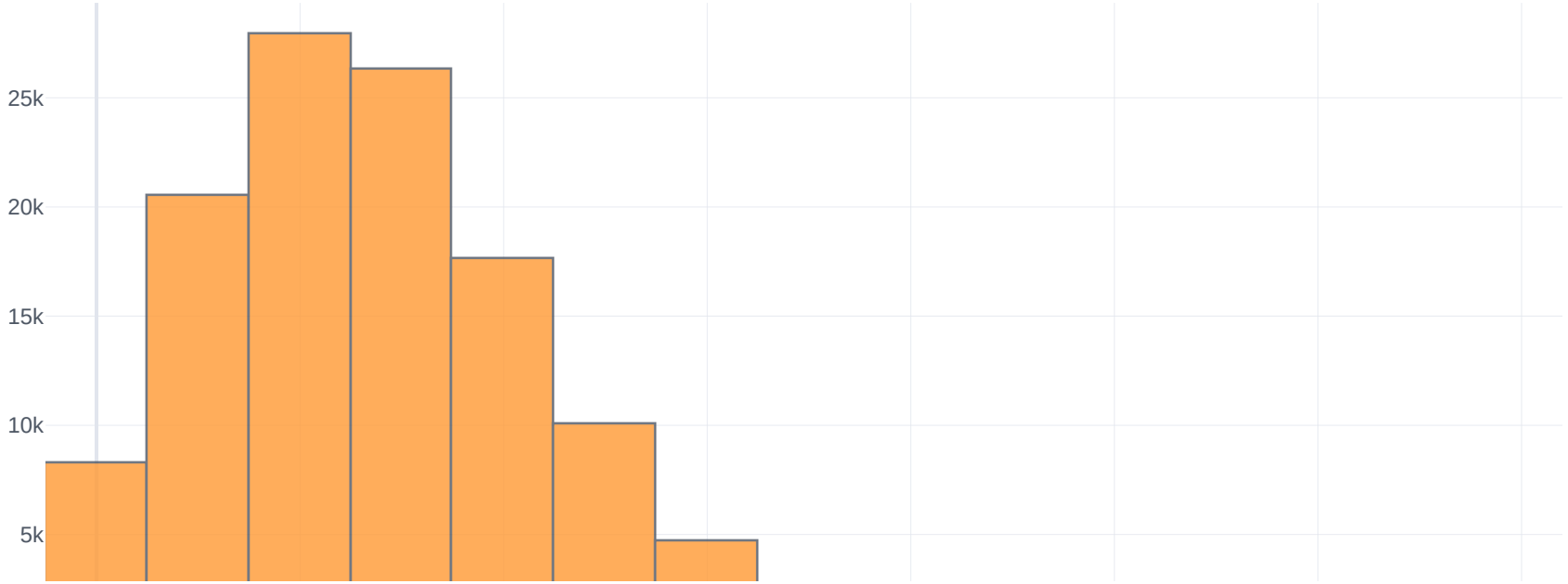
visita_minuto_primer_gol_anterior_2's histogram



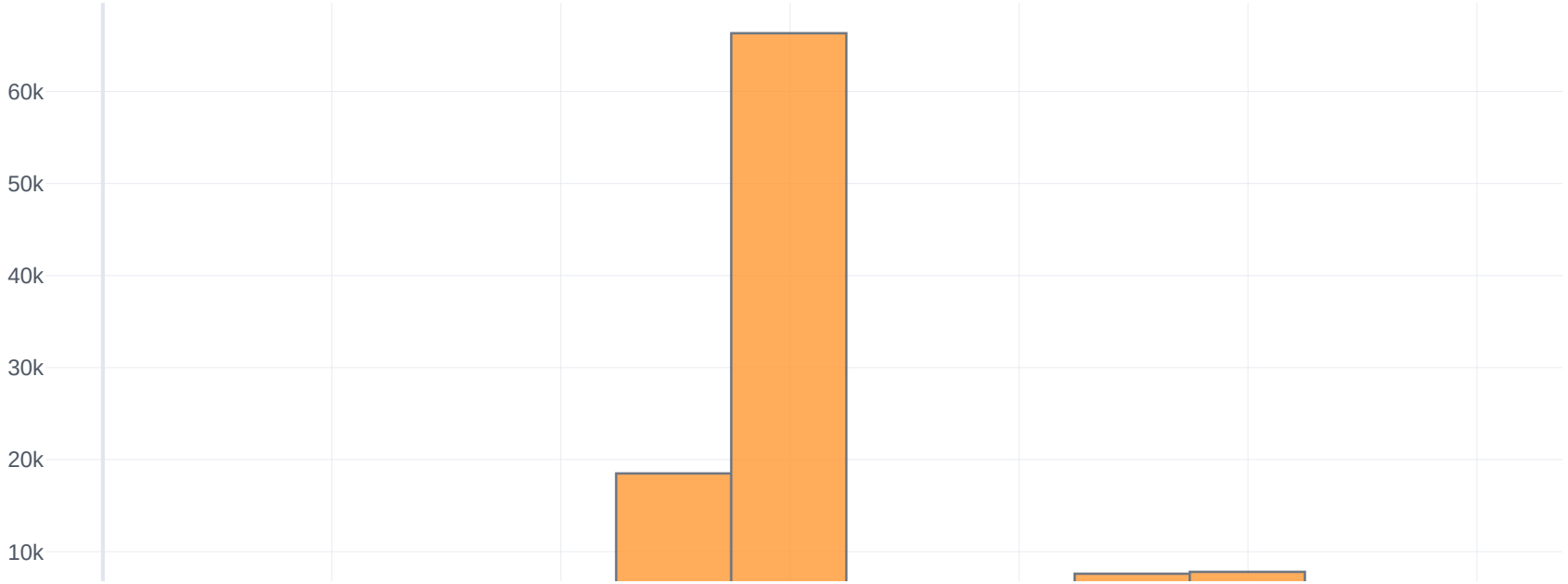
visita_minuto_ultimo_gol_anterior_2's histogram



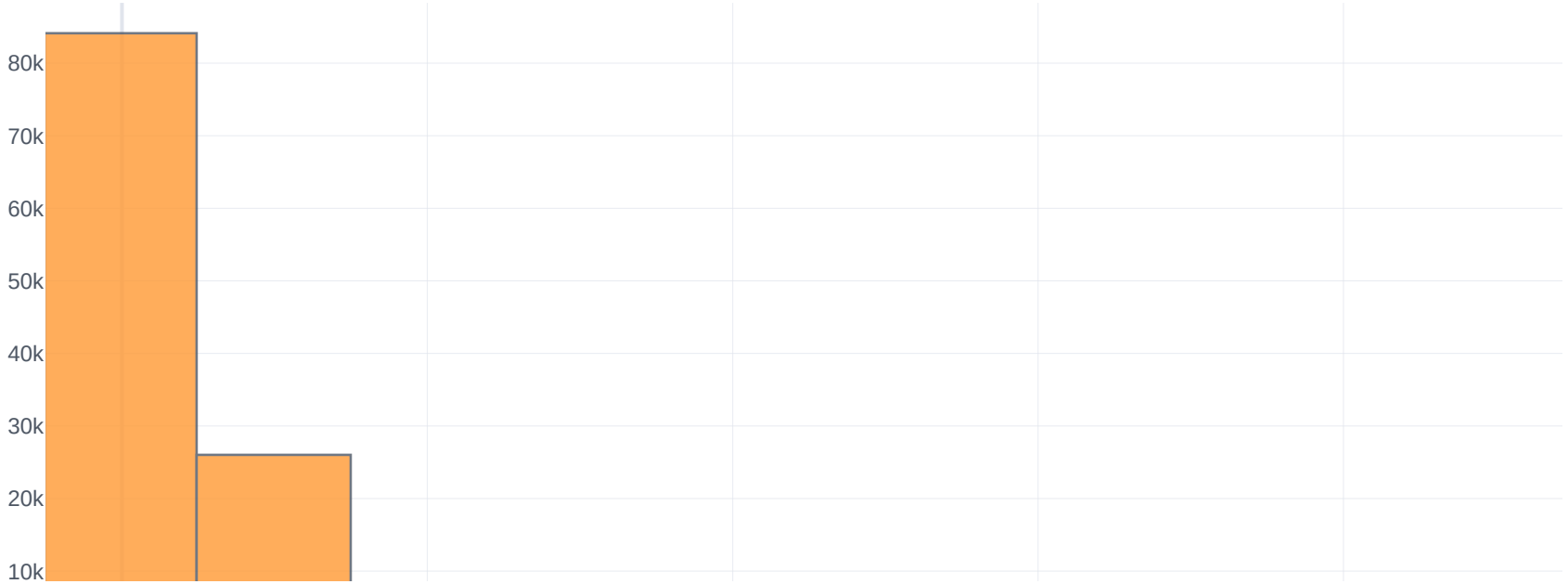
visita_number_goals_anterior_2's histogram

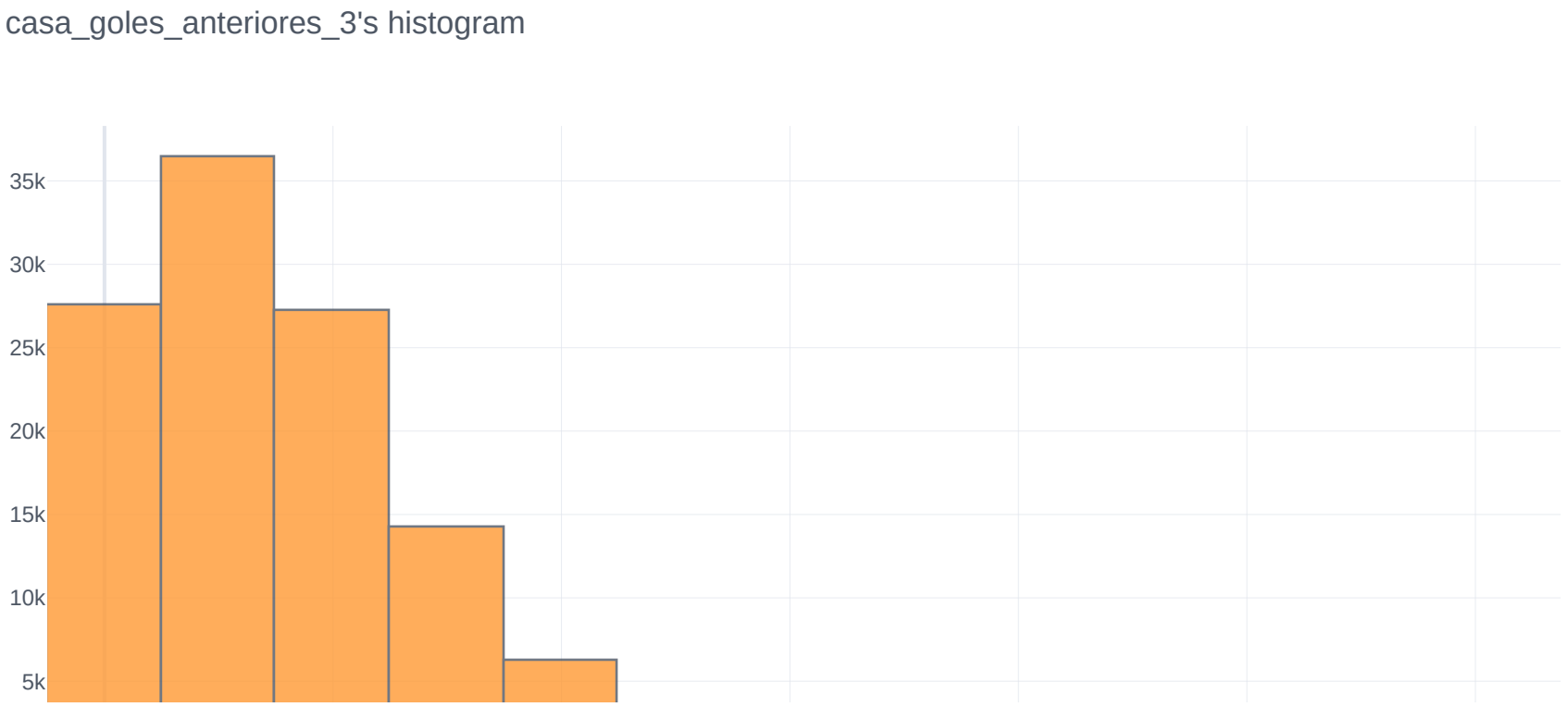


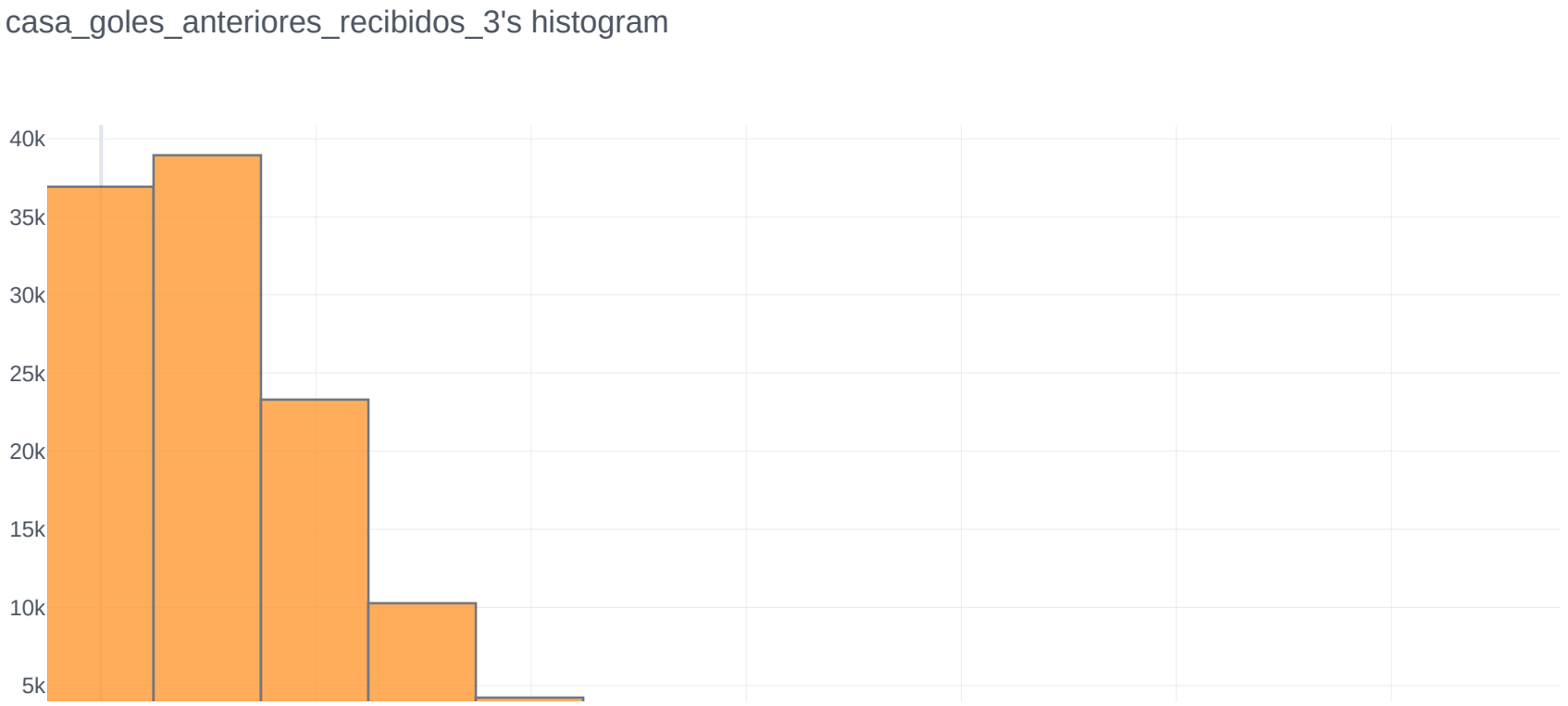
visita_sustituciones_anterior_2's histogram

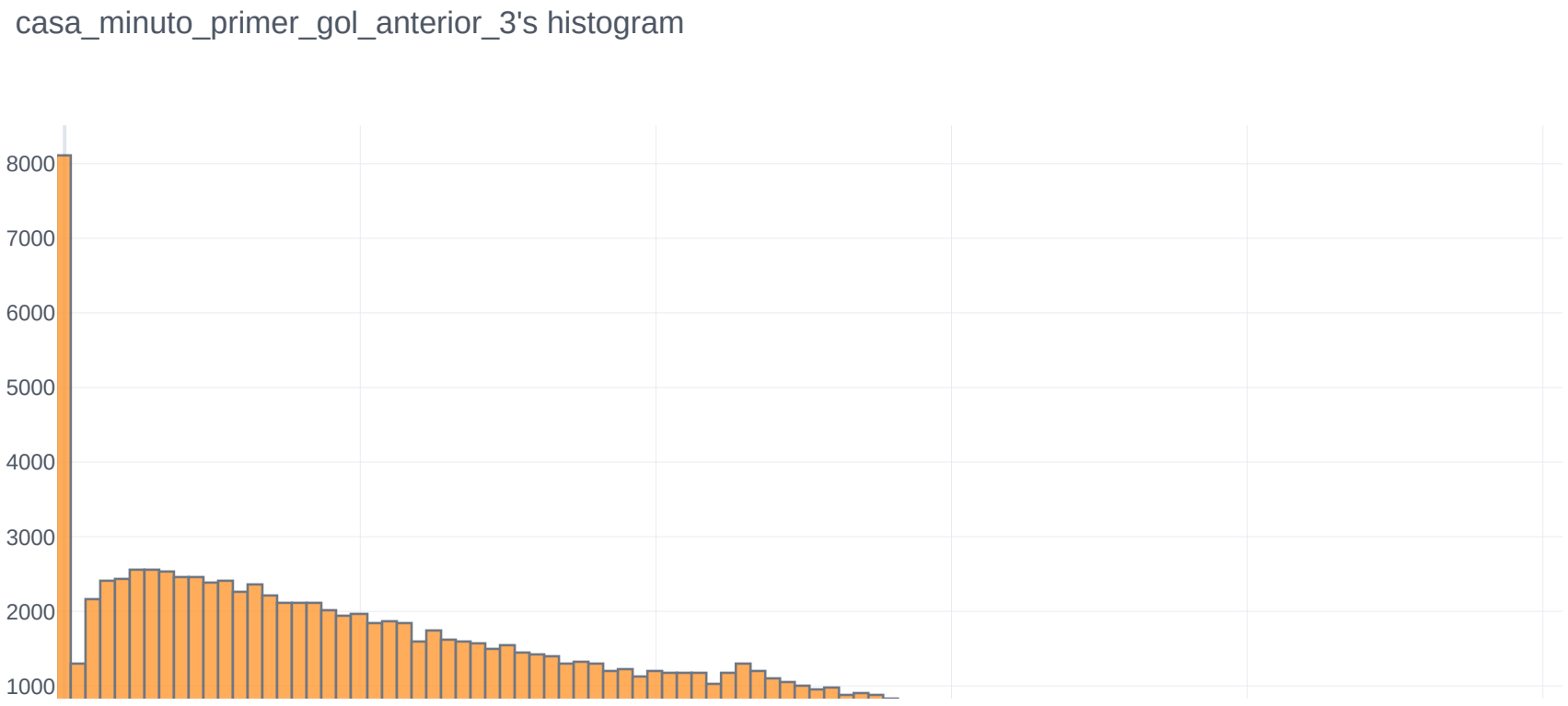


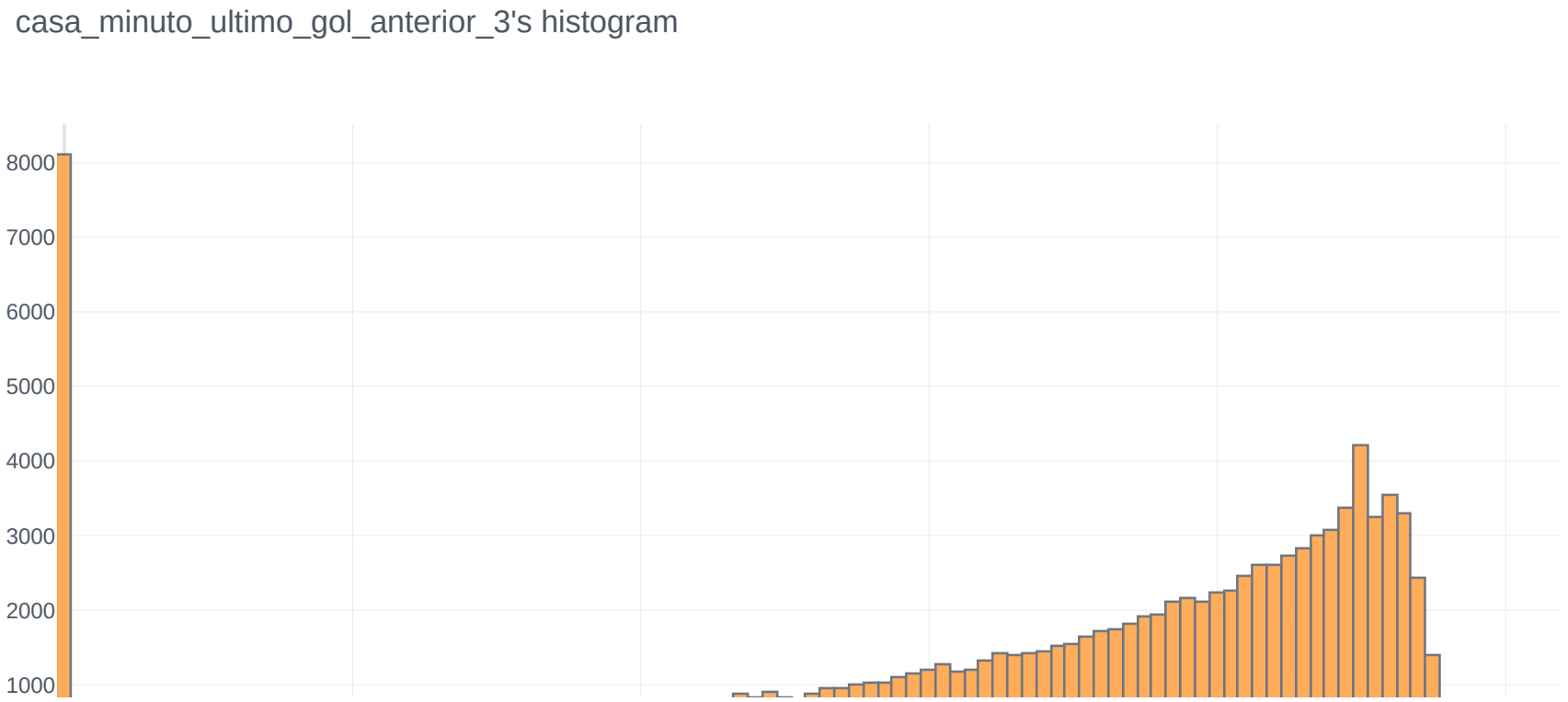
visita_goles_tiempo_añadido_anterior_2's histogram



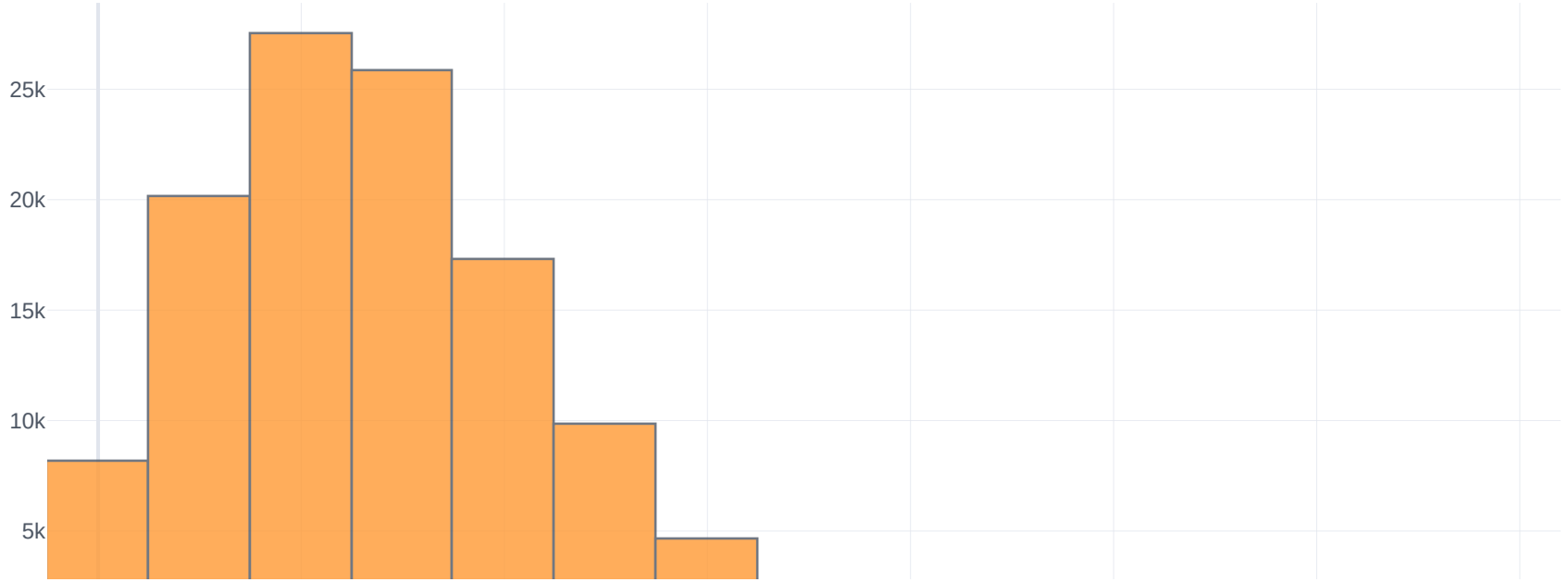

casa_goles_anteriores_3's histogram
35k
30k
25k
20k
15k
10k
5k


casa_goles_anteriores_recibidos_3's histogram
40k
35k
30k
25k
20k
15k
10k
5k

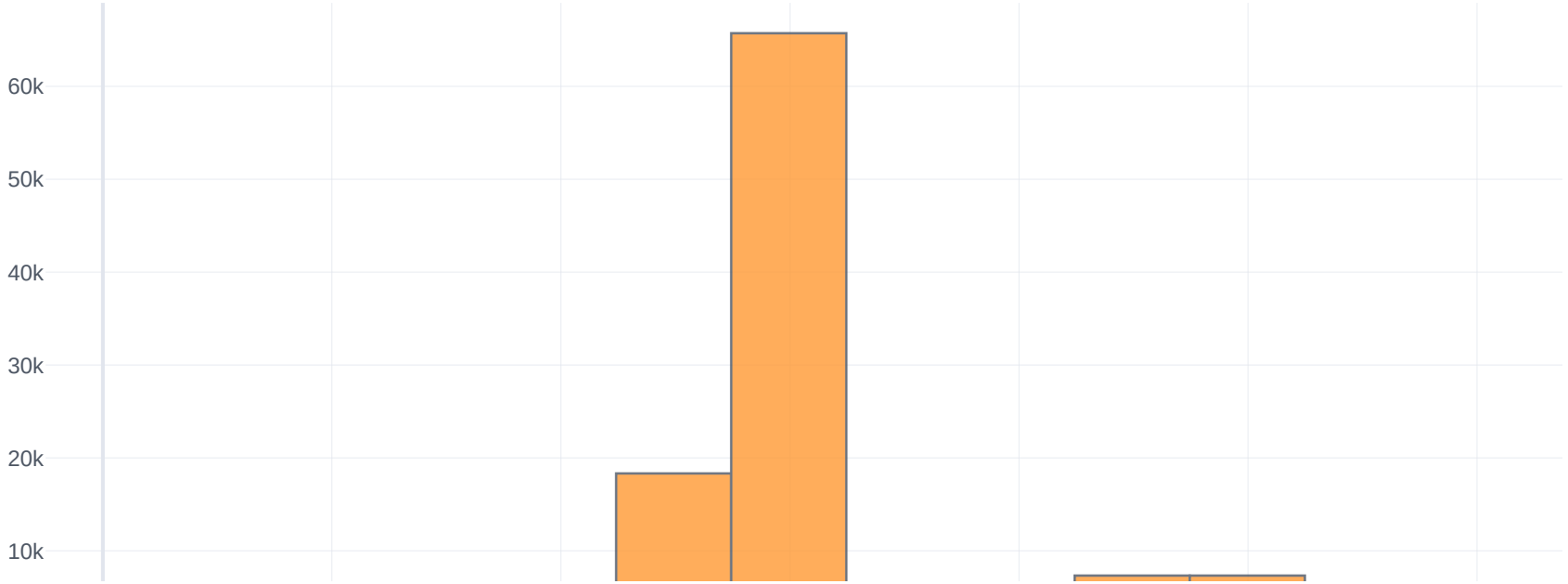

casa_minuto_primer_gol_anterior_3's histogram
8000
7000
6000
5000
4000
3000
2000
1000


casa_minuto_ultimo_gol_anterior_3's histogram
8000
7000
6000
5000
4000
3000
2000
1000

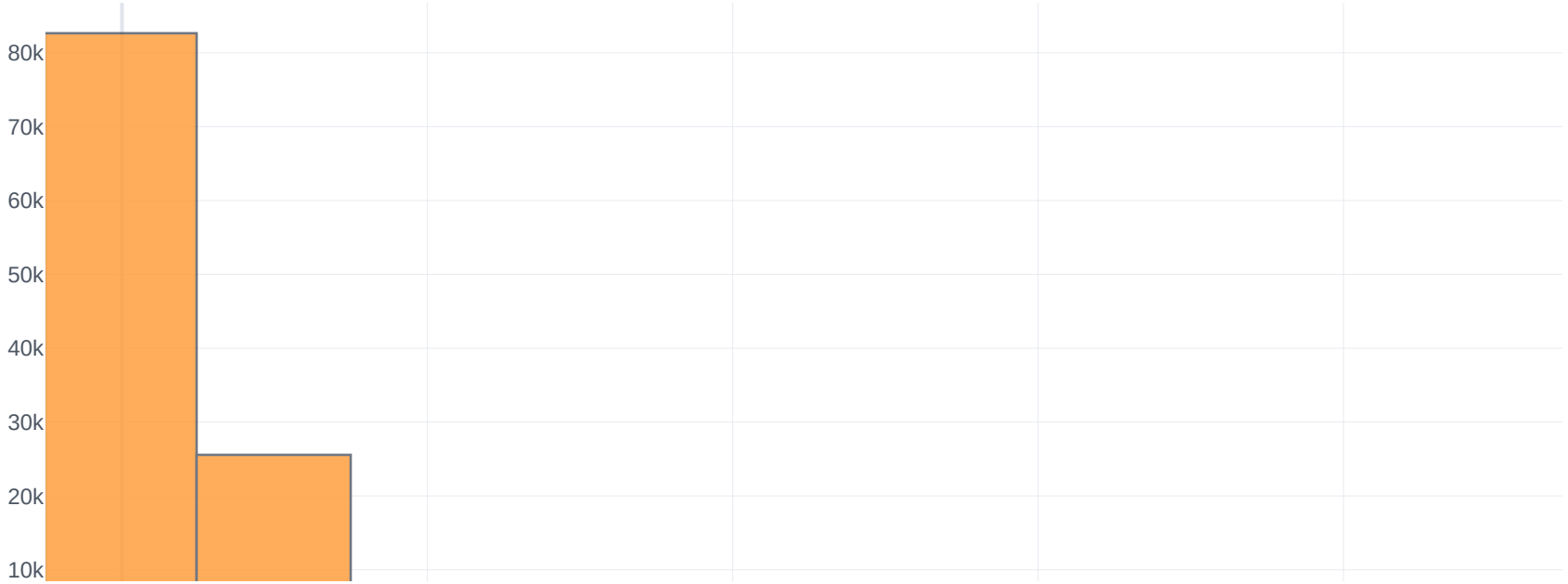
casa_number_goals_anterior_3's histogram



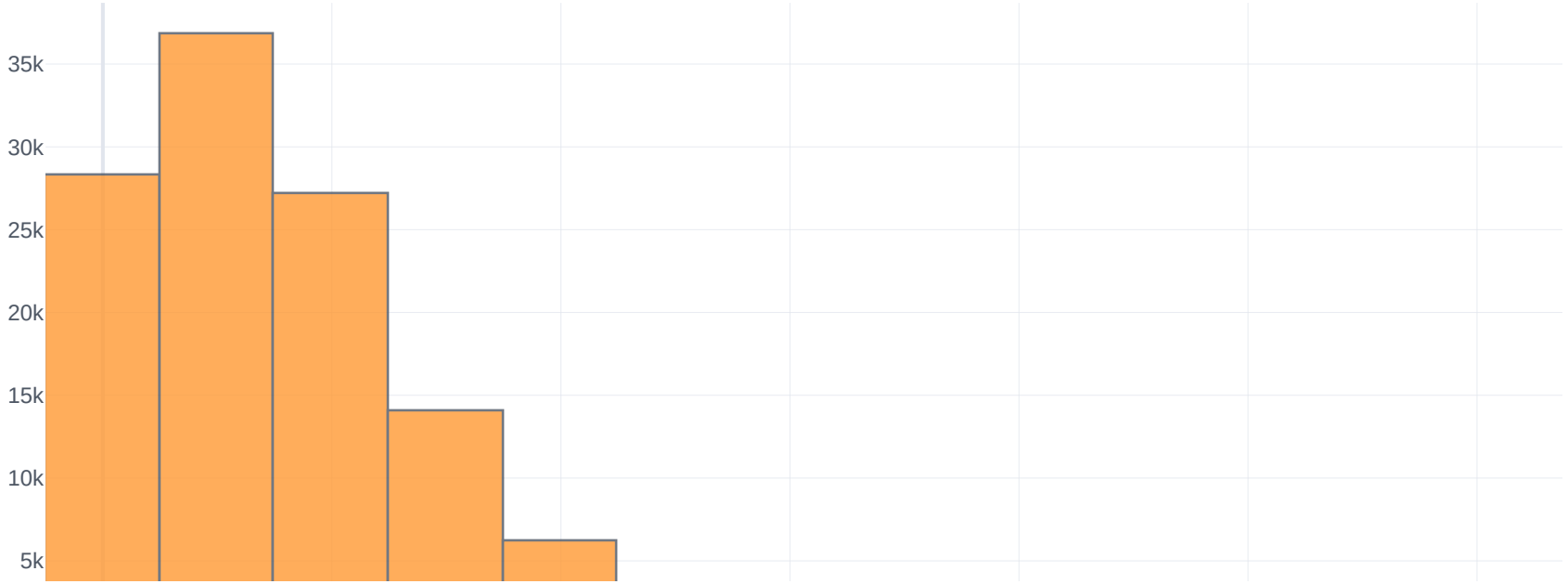
casa_sustituciones_anterior_3's histogram



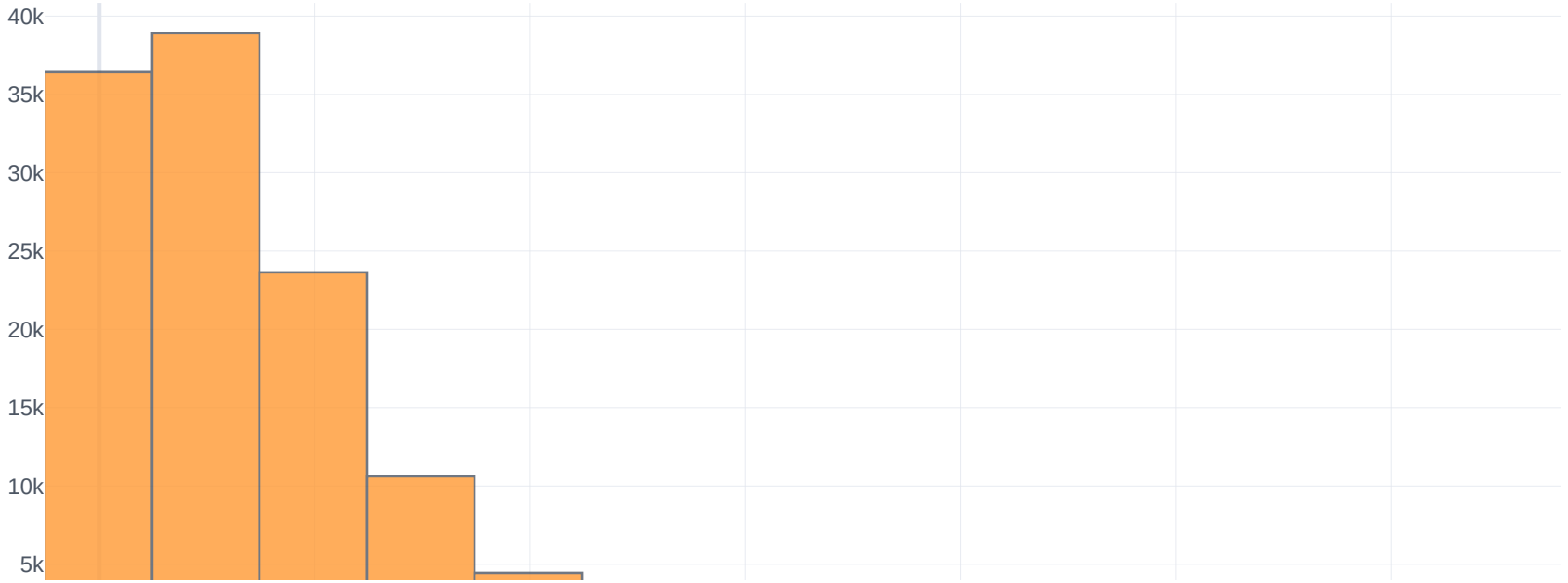
casa_goles_tiempo_añadido_anterior_3's histogram



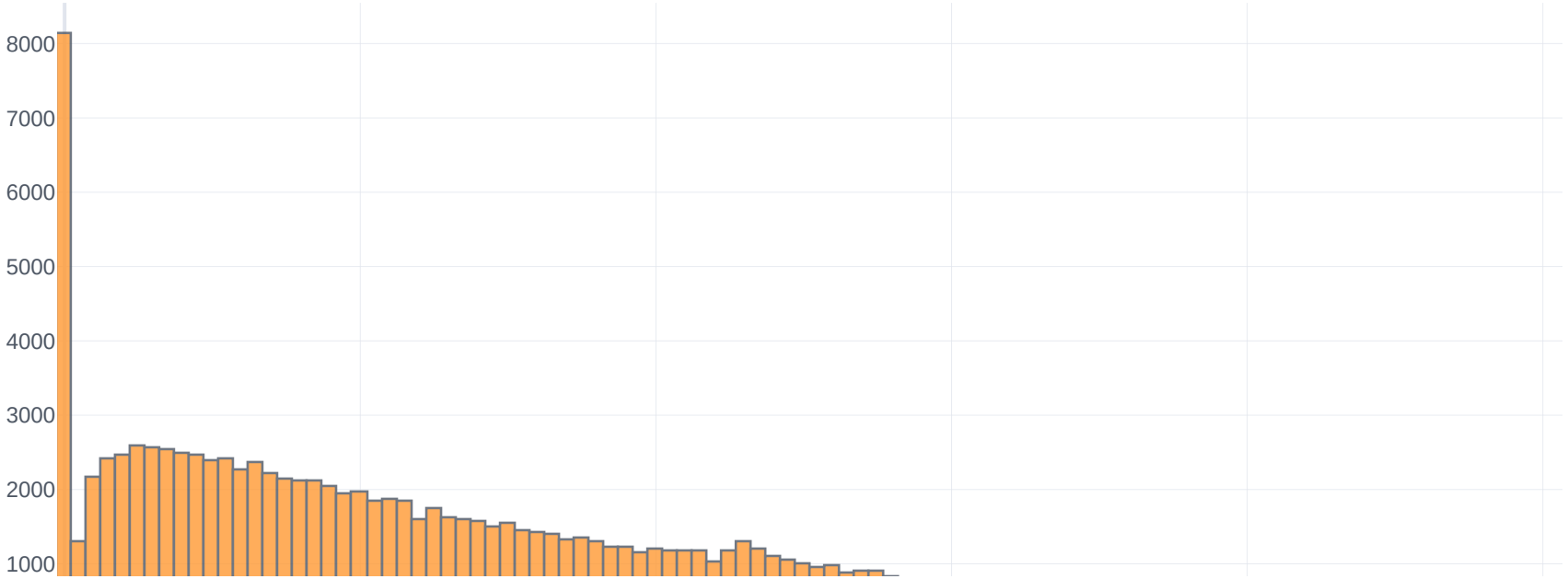
visita_goles_anteriores_recibidos_3's histogram



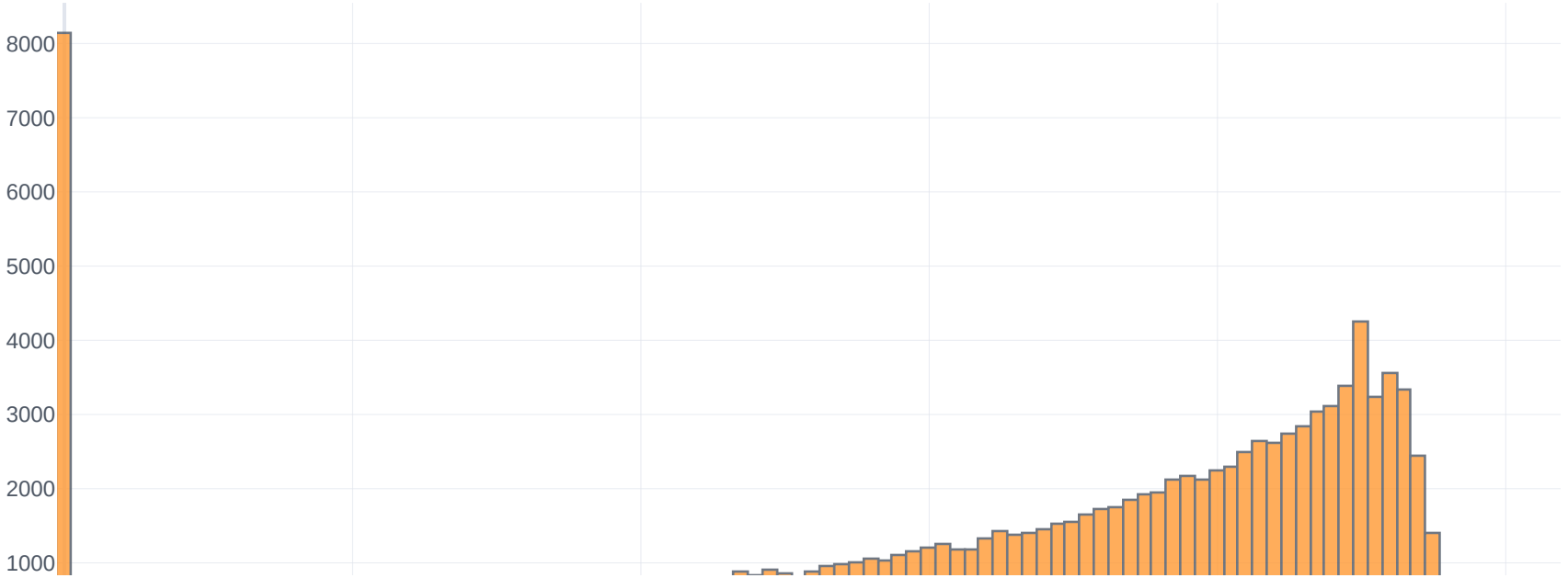
visita_goles_anteriores_3's histogram



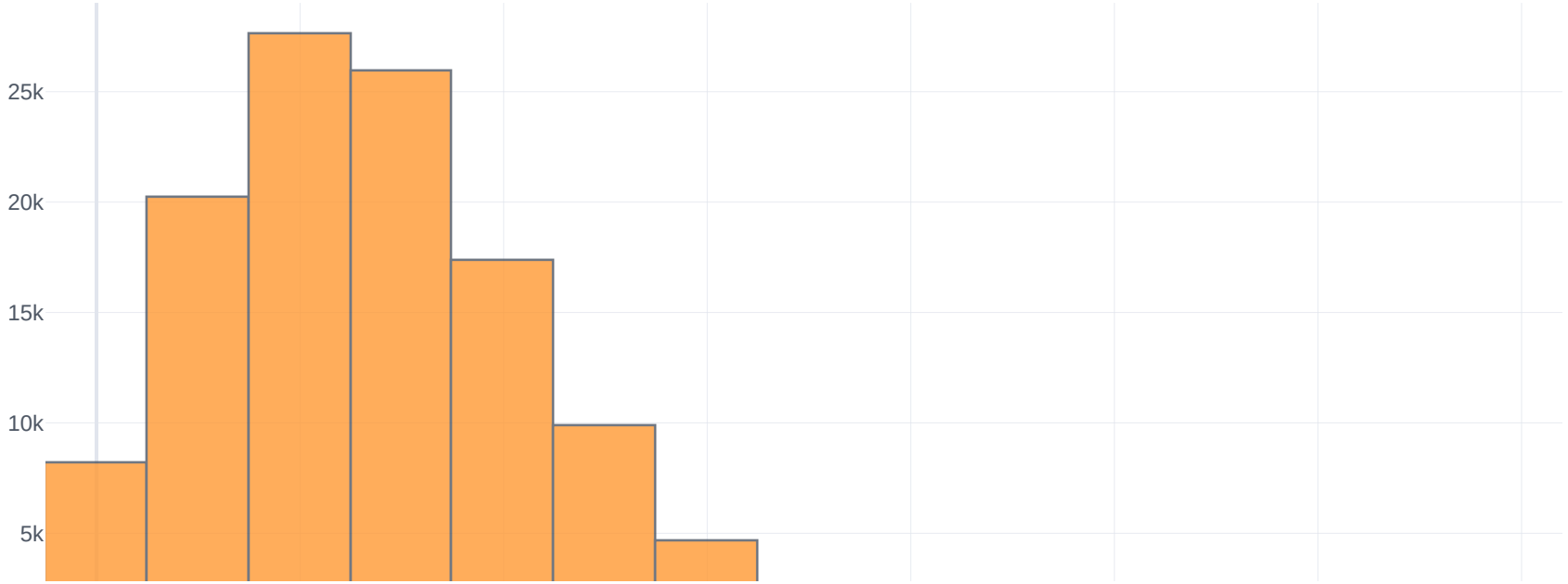
visita_minuto_primer_gol_anterior_3's histogram



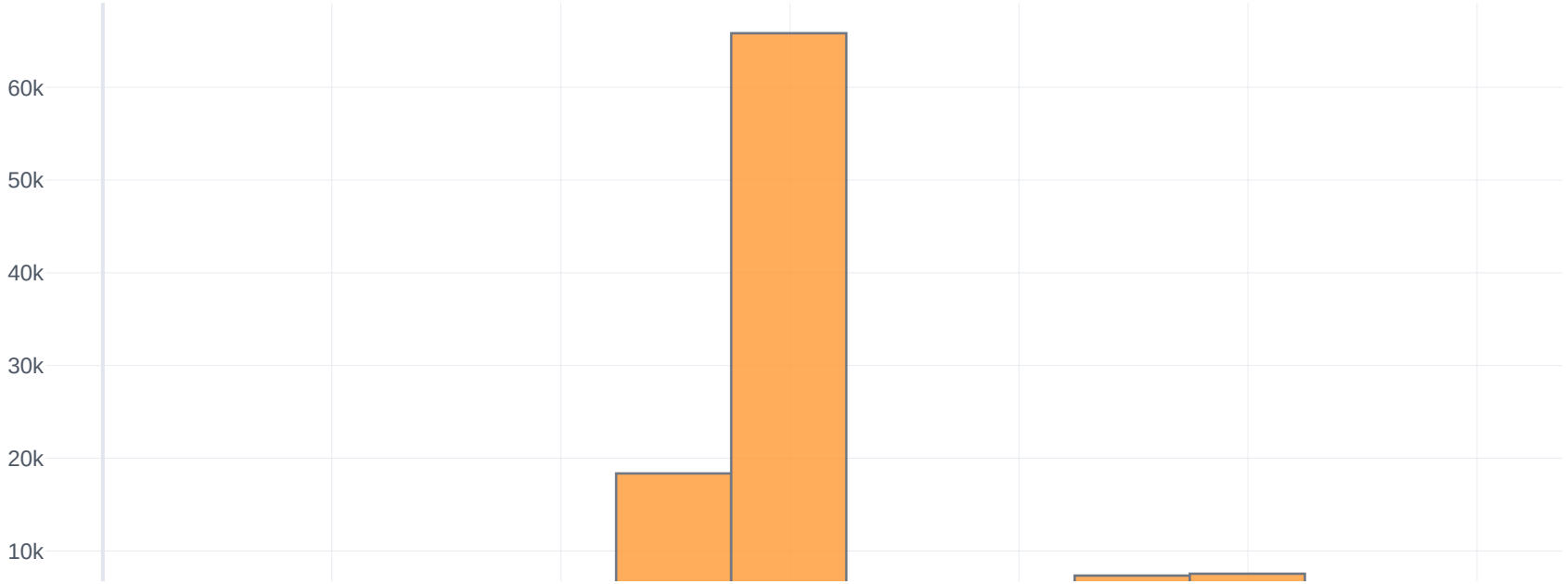
visita_minuto_ultimo_gol_anterior_3's histogram



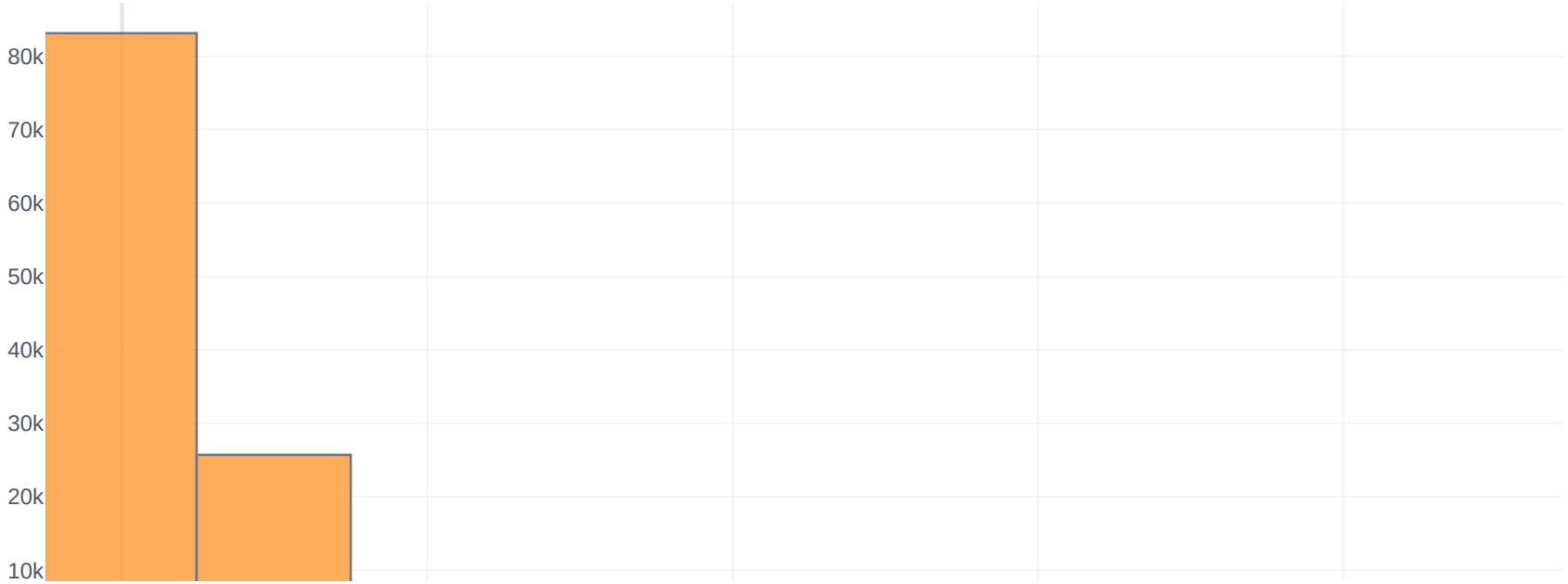
visita_number_goals_anterior_3's histogram



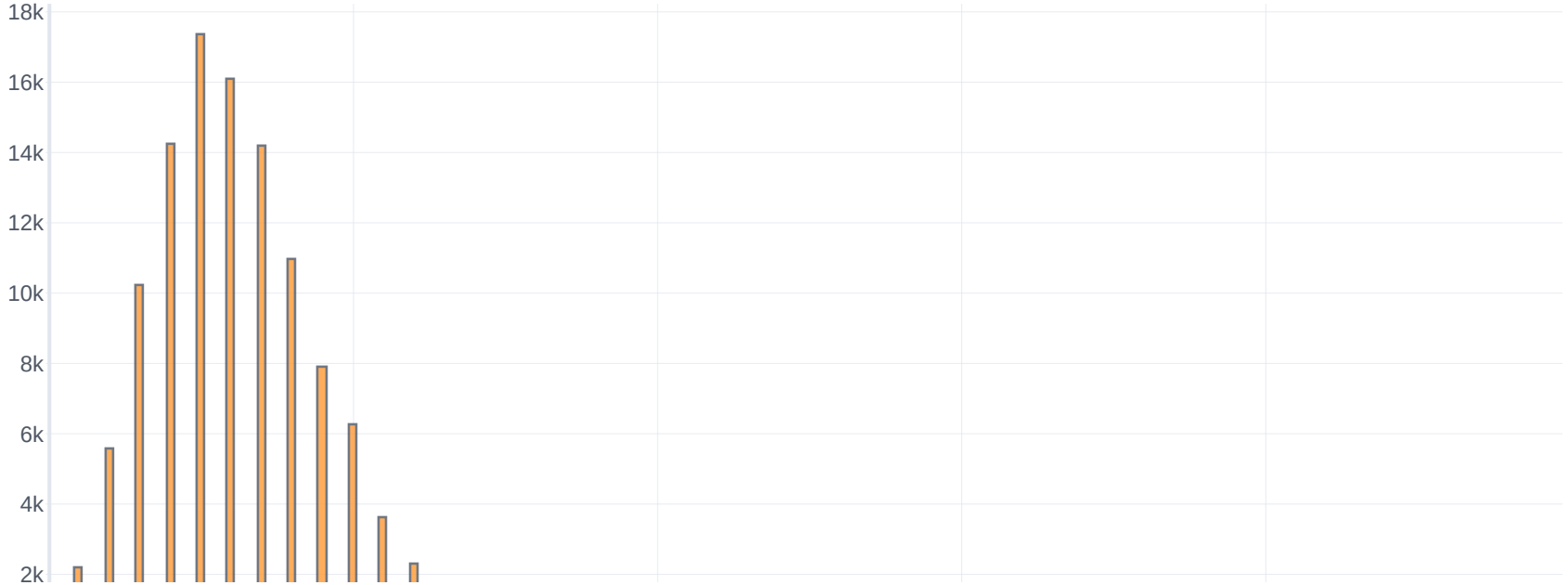
visita_sustituciones_anterior_3's histogram

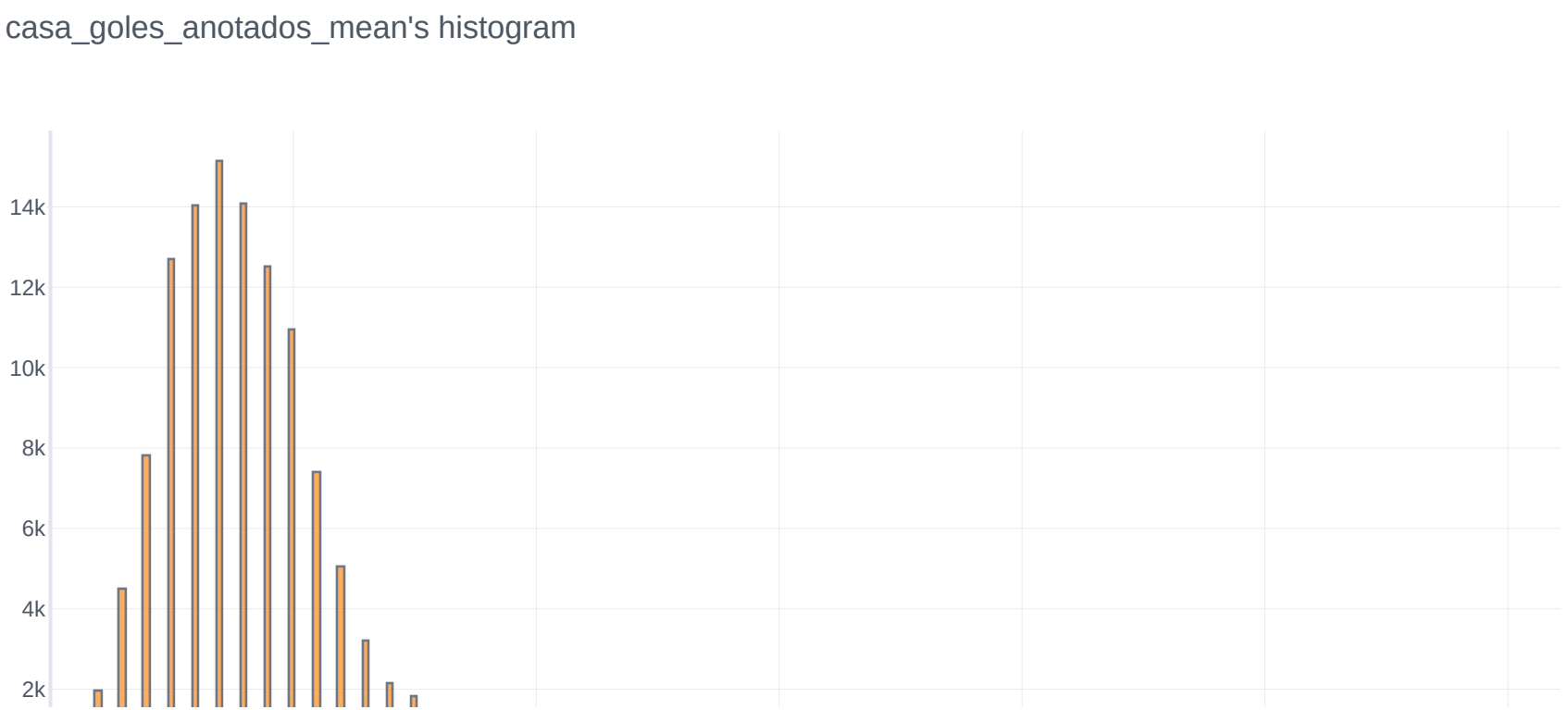


visita_goles_tiempo_añadido_anterior_3's histogram

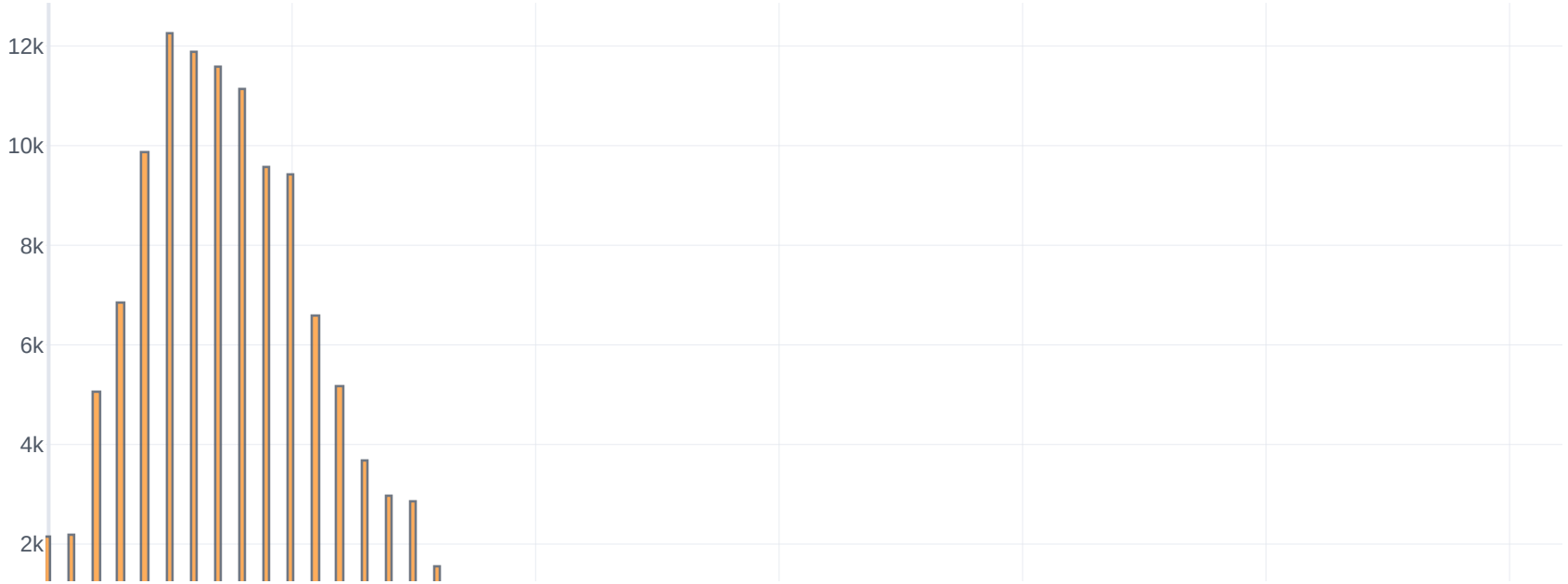


casa_goles_recibidos_mean's histogram



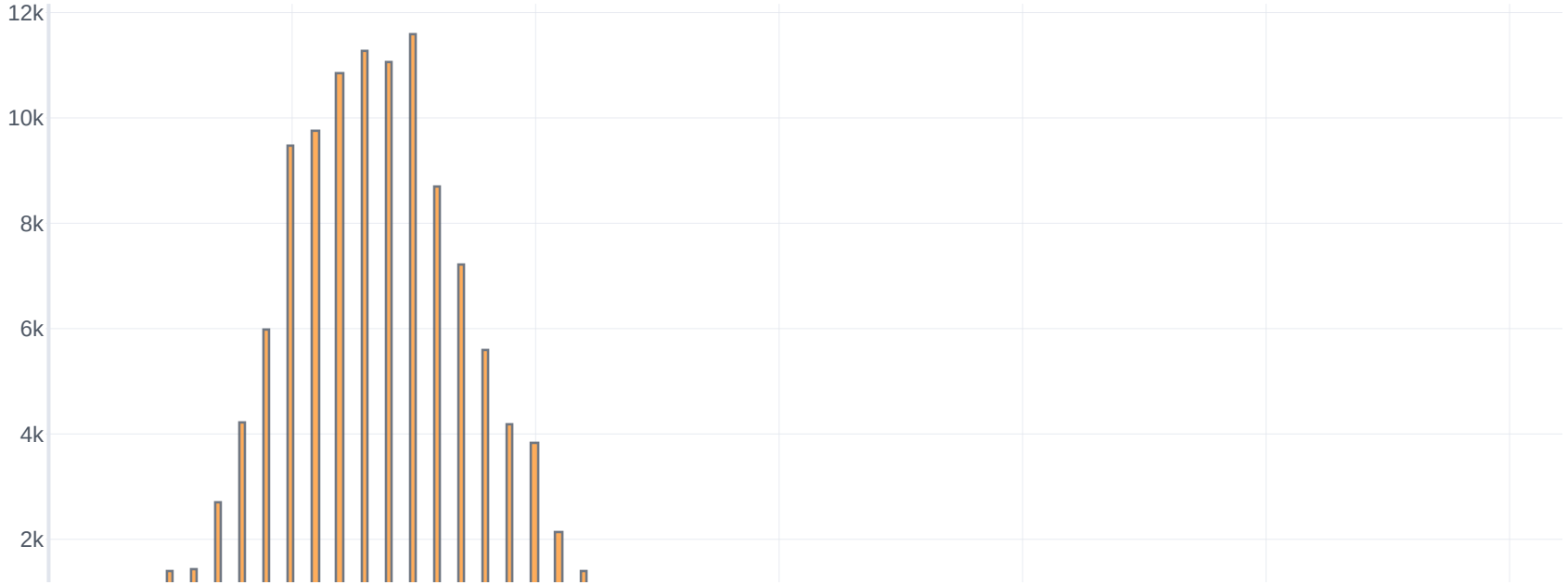

casa_goles_anotados_mean's histogram
14k
12k
10k
8k
6k
4k
2k

visita_goles_recibidos_mean's histogram

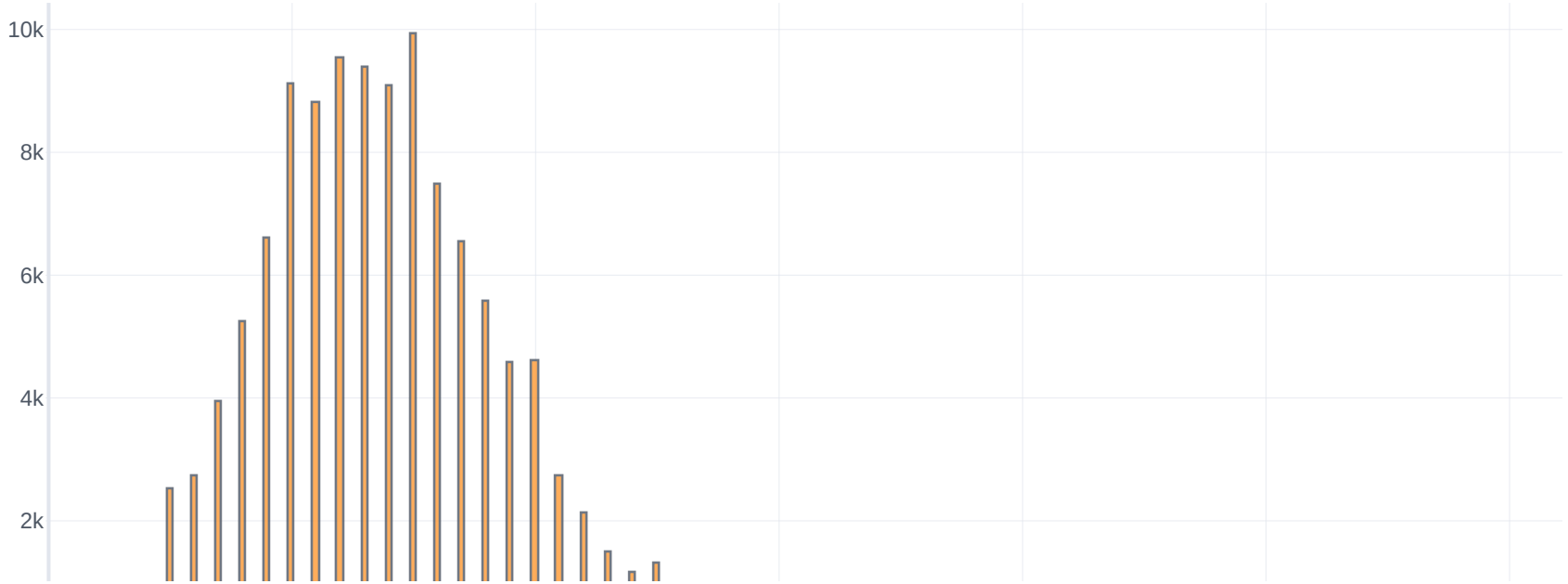


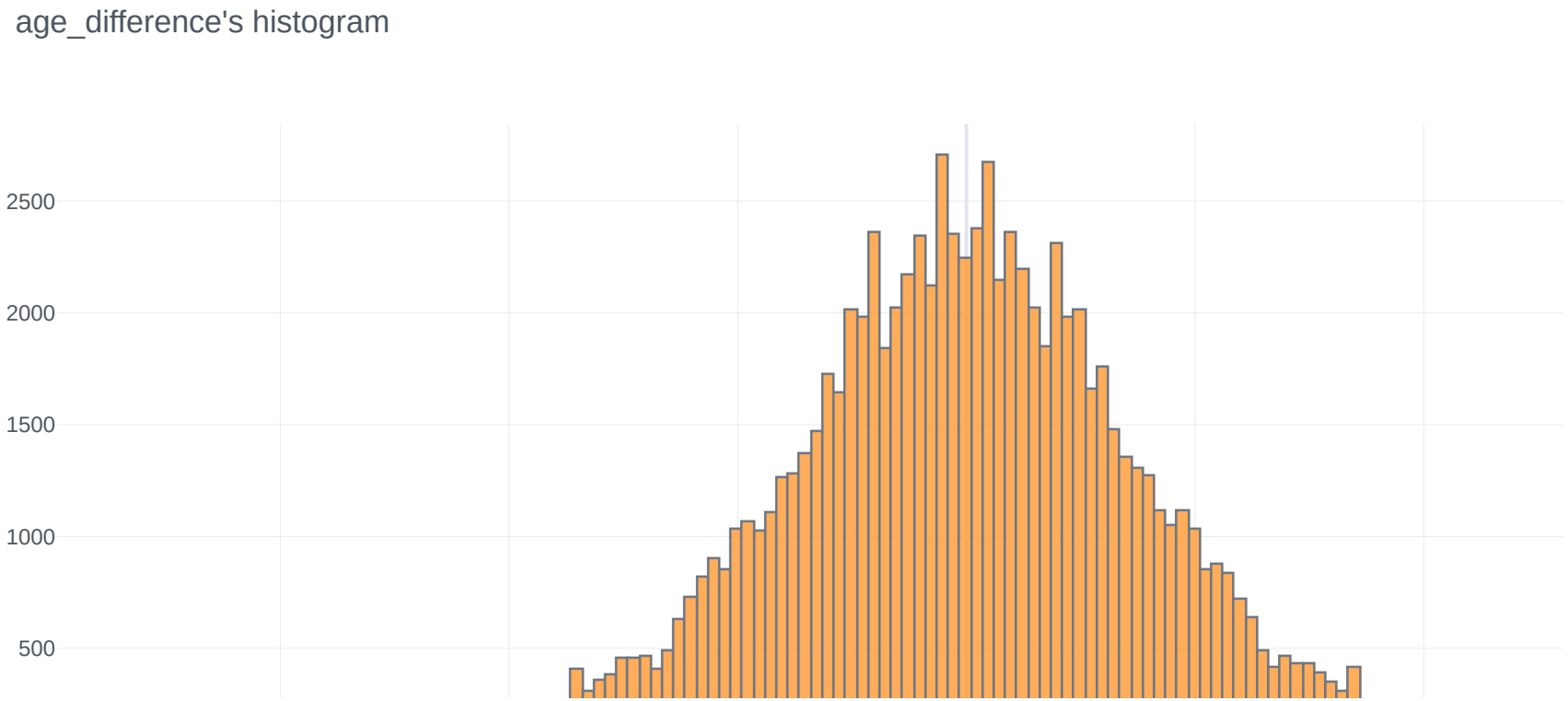
visita_goles_anotados_mean's histogram

casa_number_goals_mean's histogram

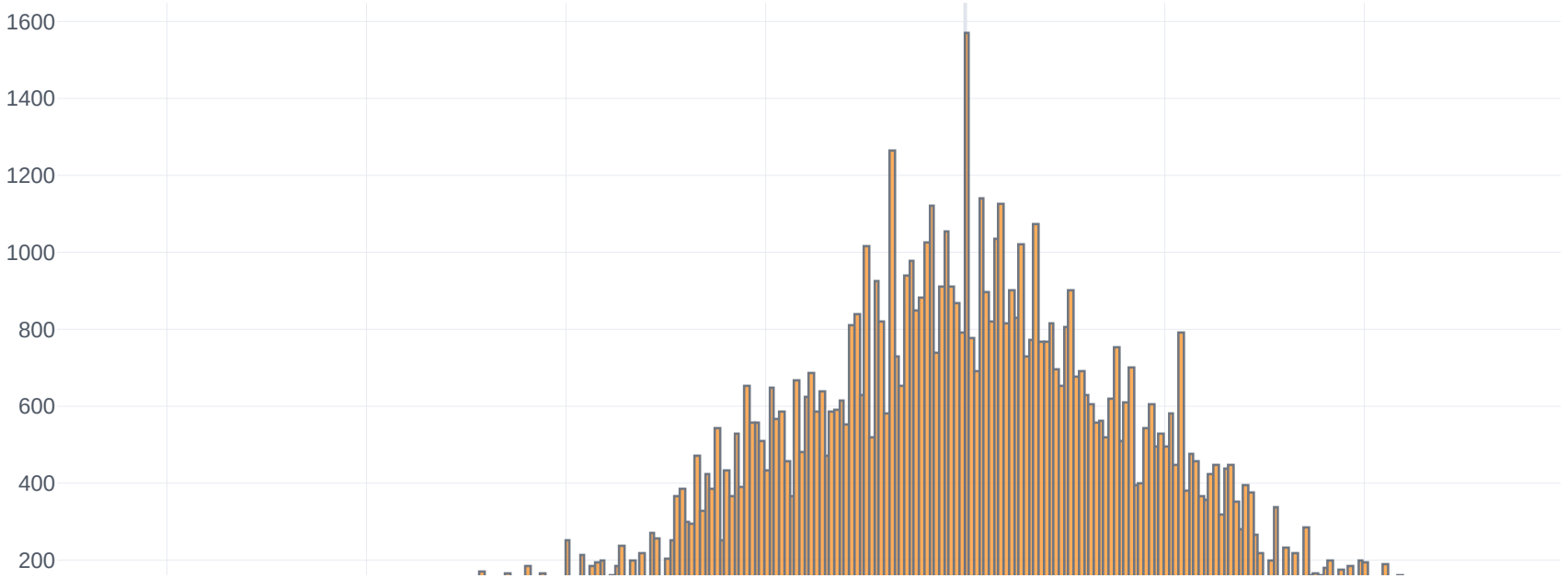


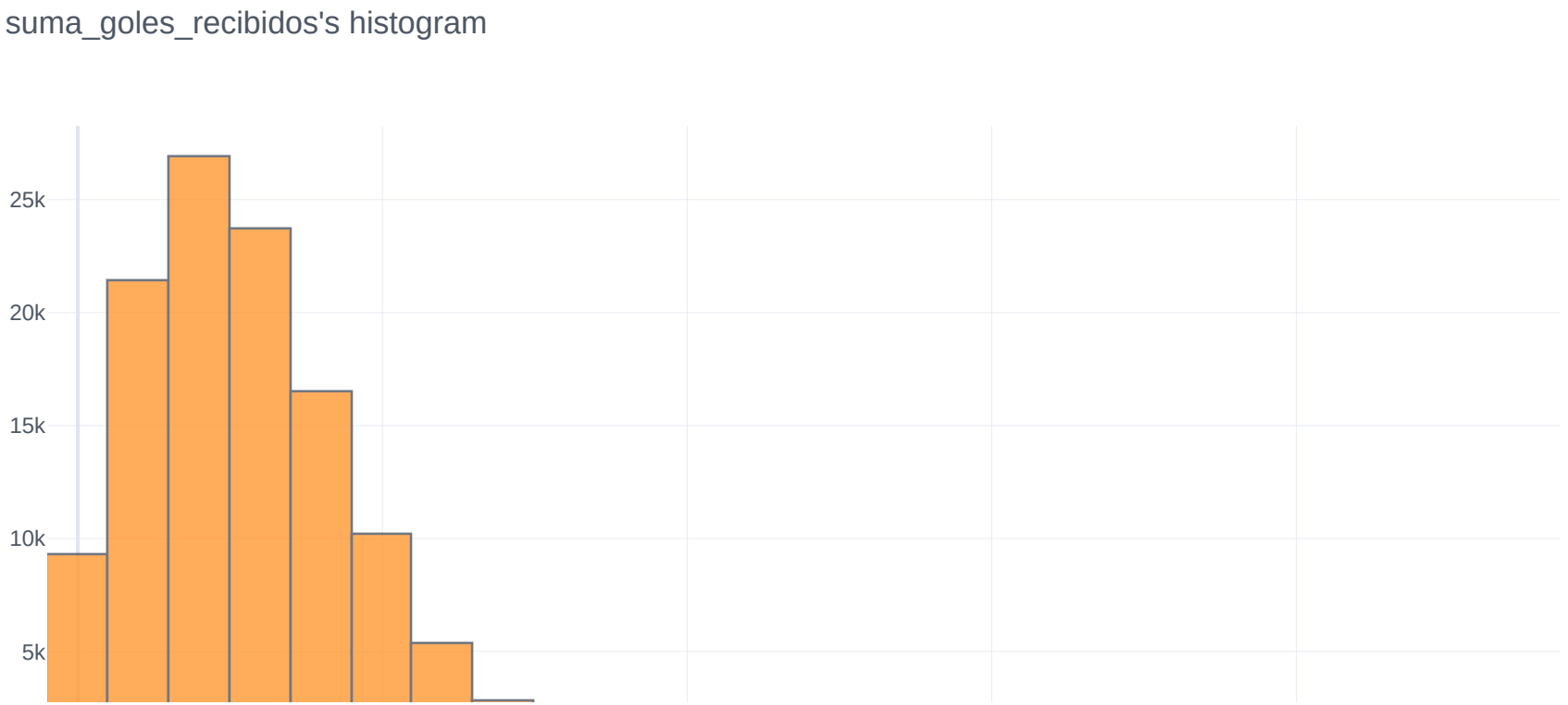
visita_number_goals_mean's histogram



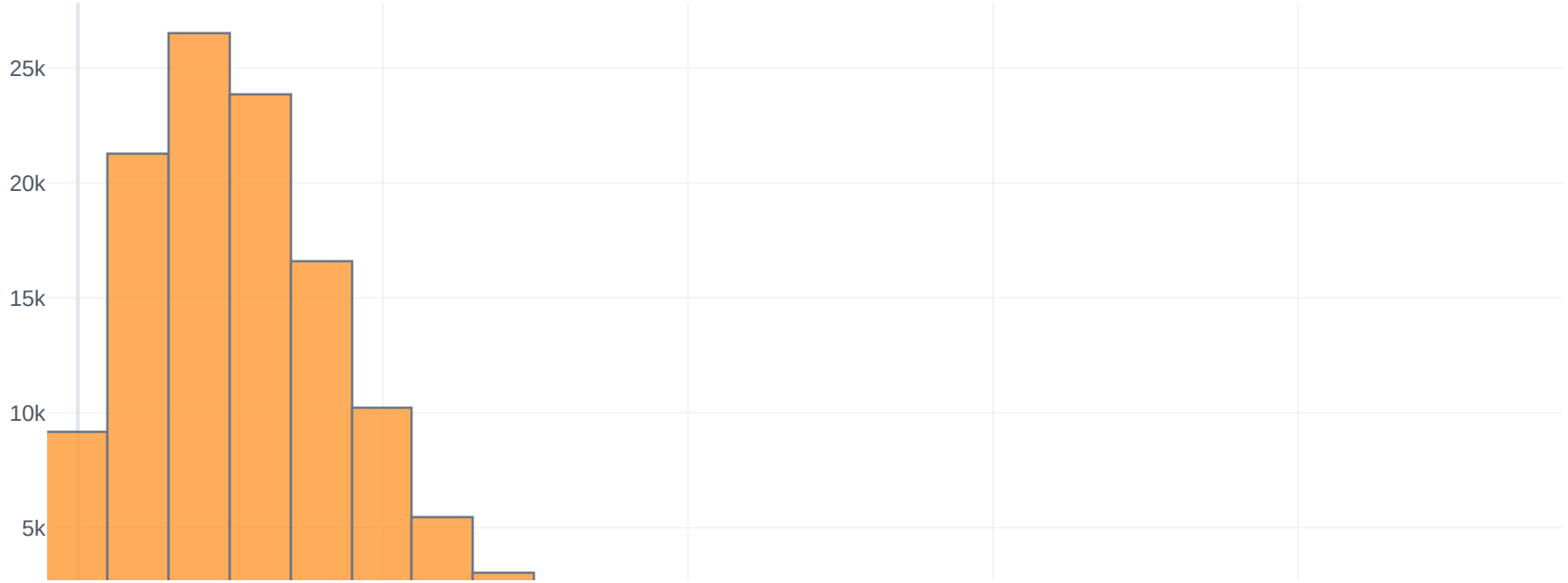

age_difference's histogram
2500
2000
1500
1000
500

foreigners_percentage_difference's histogram

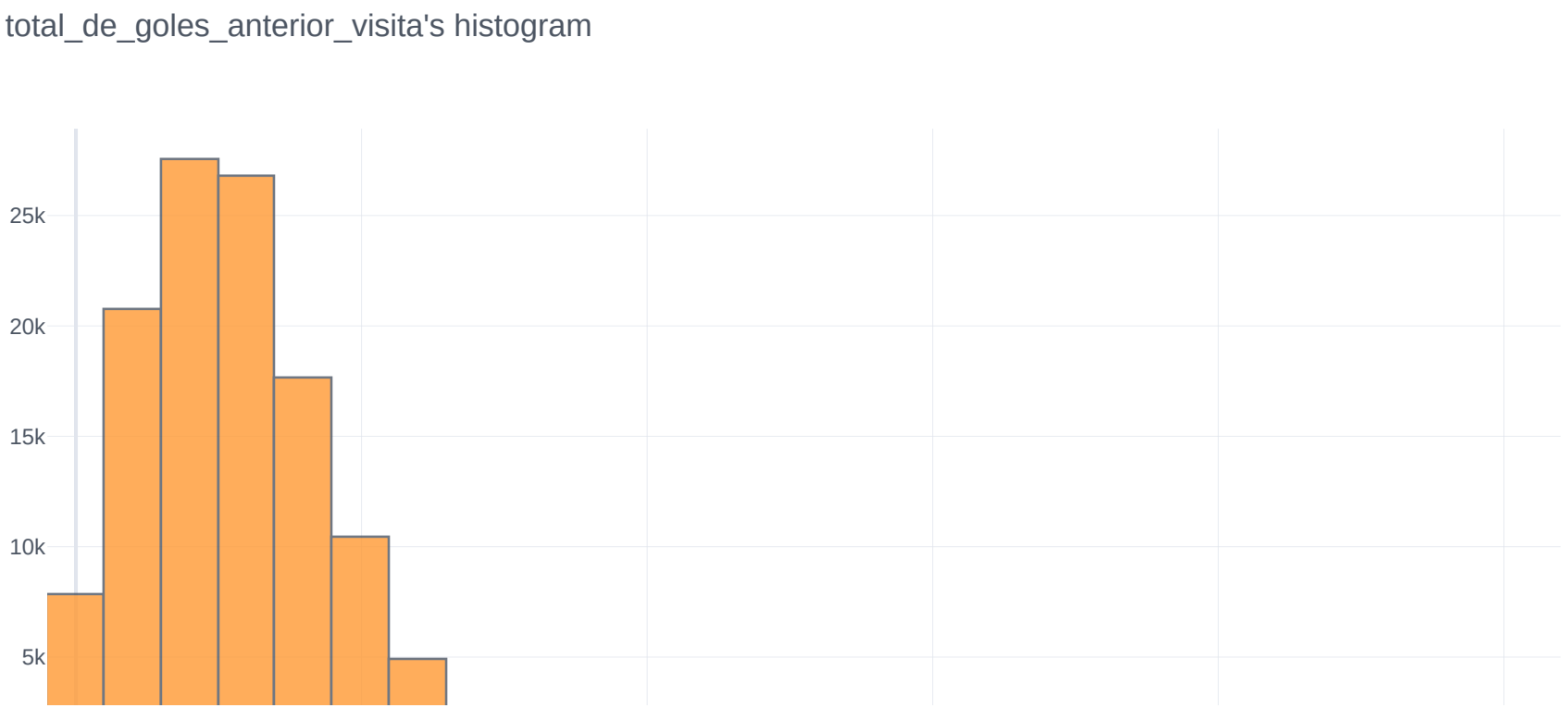



suma_goles_recibidos's histogram
25k
20k
15k
10k
5k

suma_goles_anotados's histogram

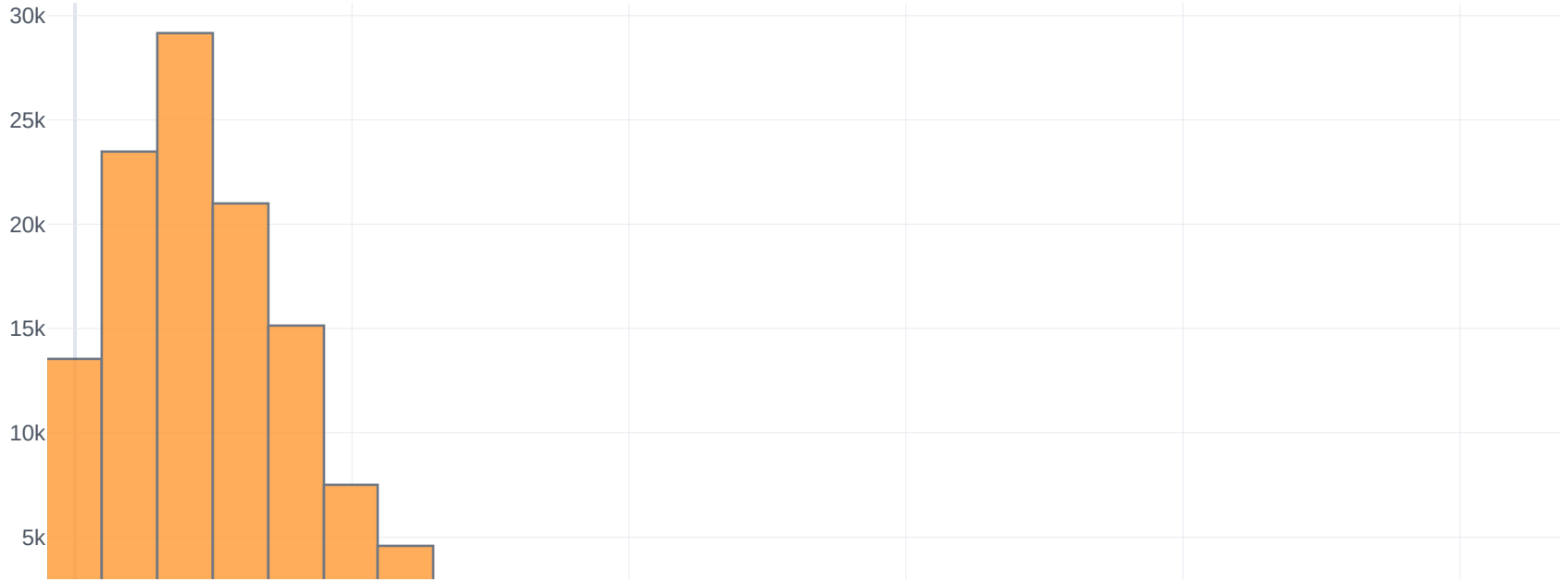



total_de_goles_anterior_casa's histogram
25k
20k
15k
10k
5k

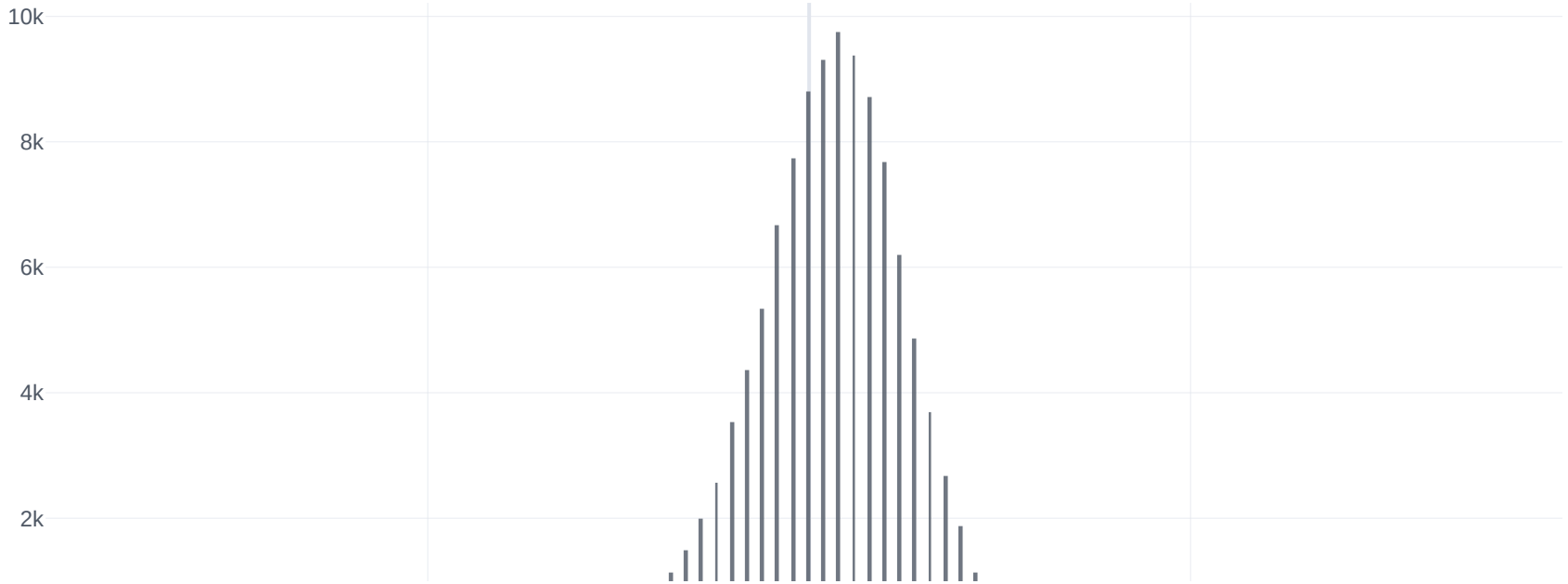

total_de_goles_anterior_visita's histogram
25k
20k
15k
10k
5k

anotado_local_recibido_visita's histogram

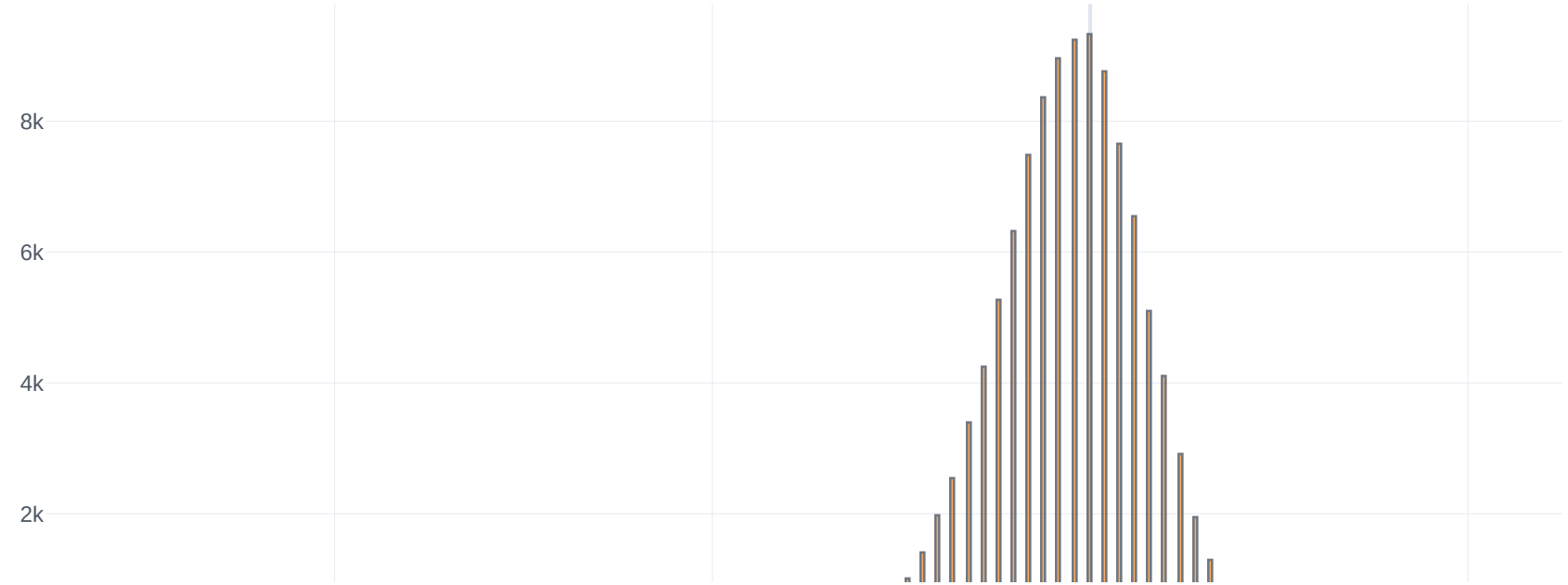
anotado_visita_recibido_local's histogram



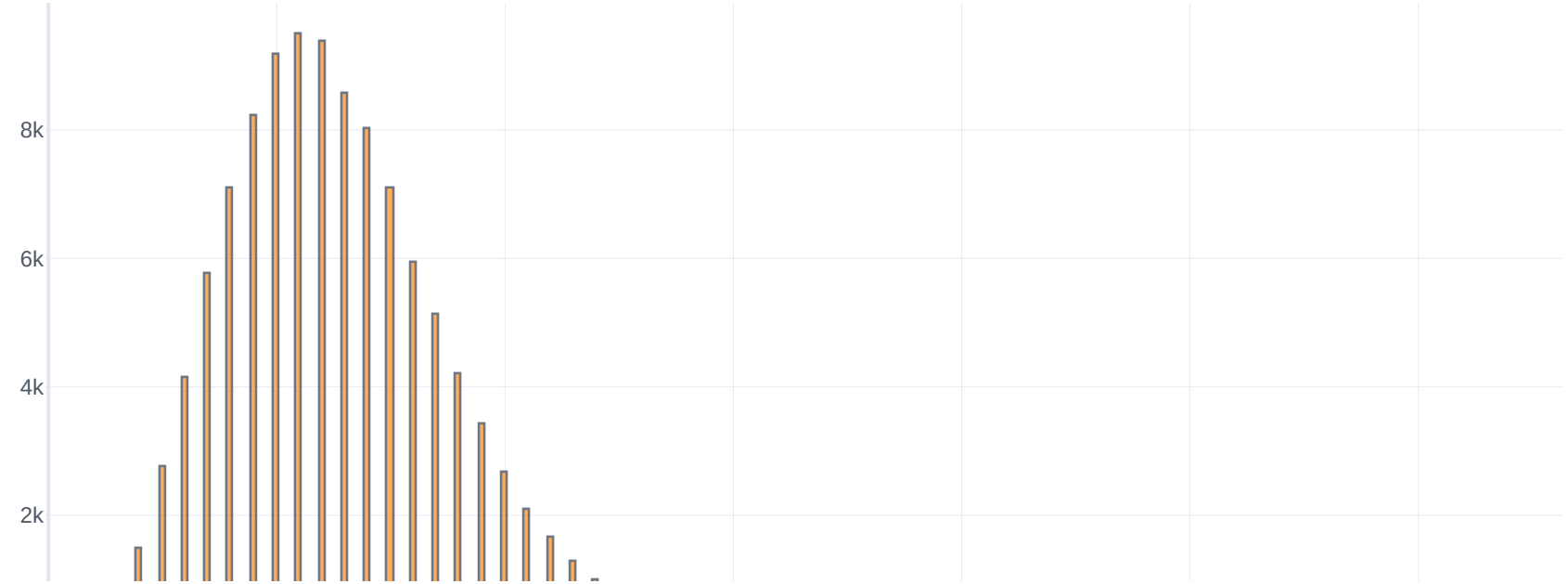
diferencia_promedio_anotados's histogram



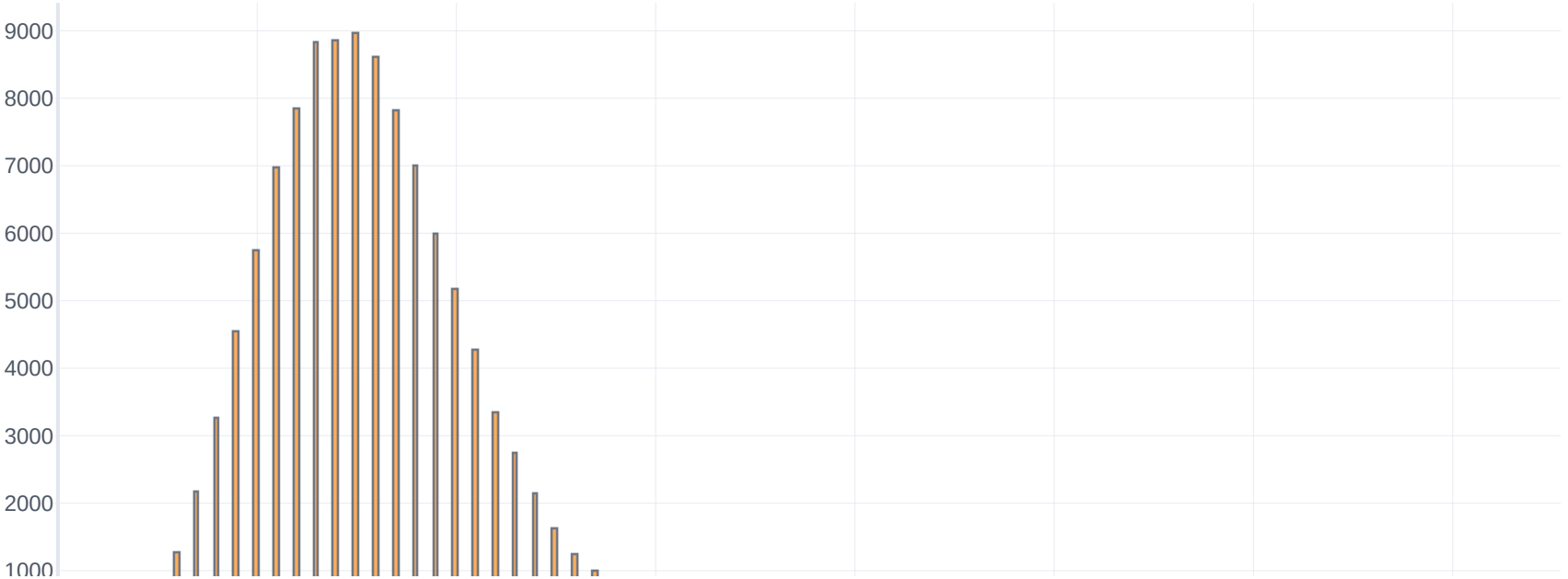
diferencia_promedio_recibidos's histogram

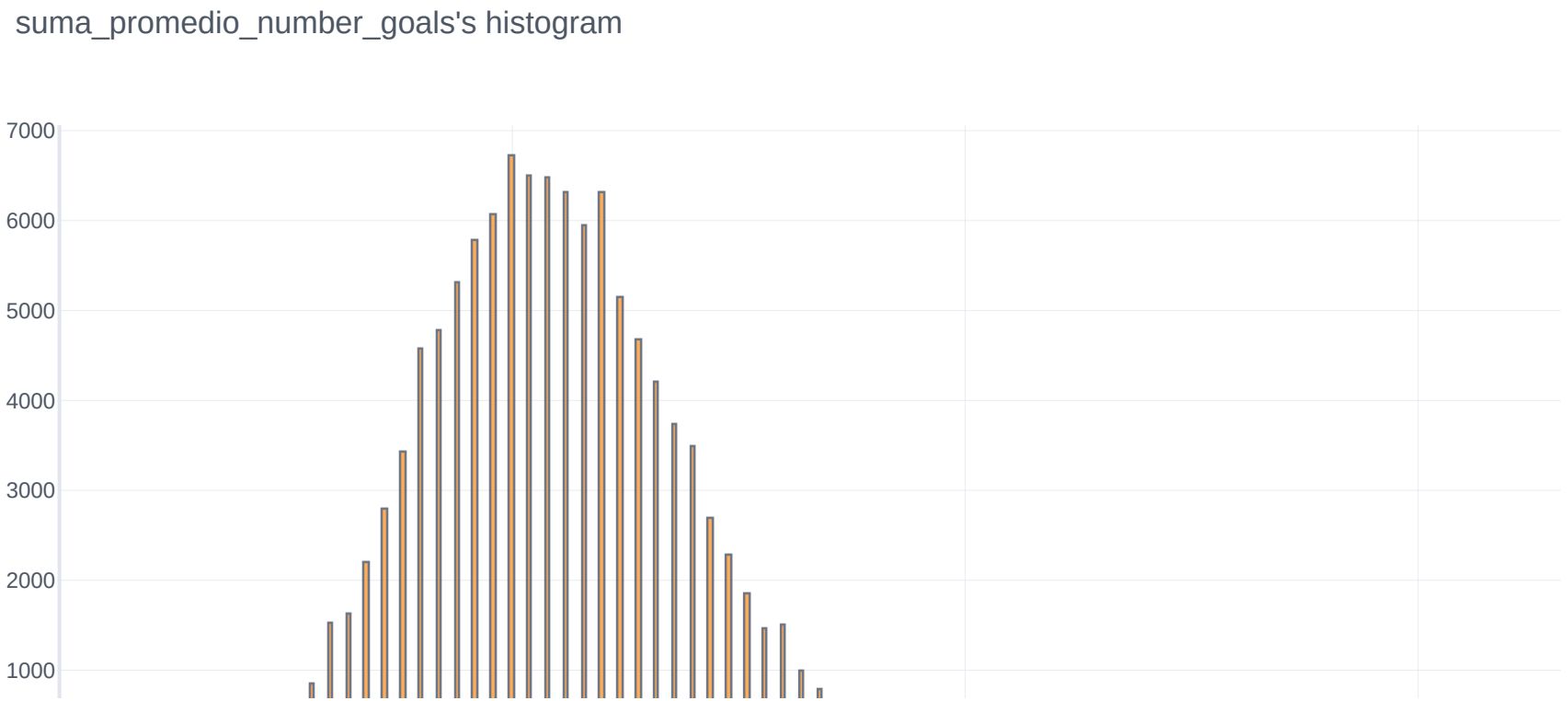


suma_promedio_recibidos_anotados's histogram

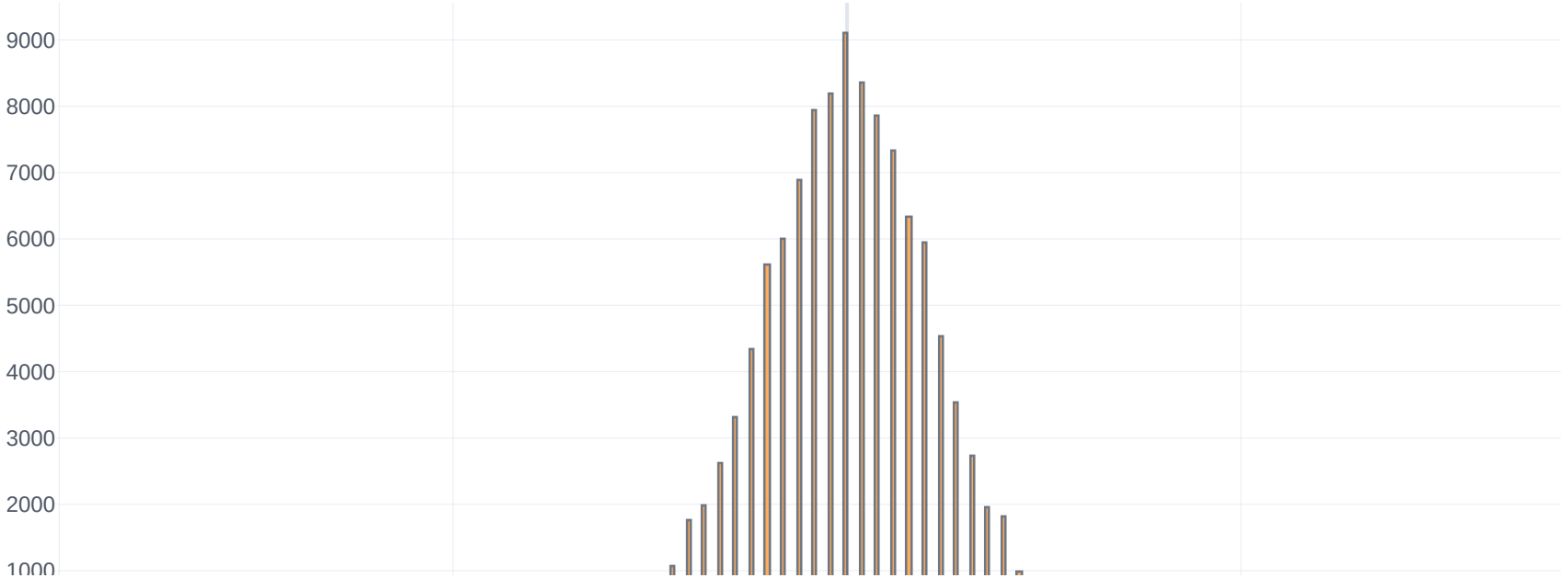


suma_promedio_anotados_recibidos's histogram

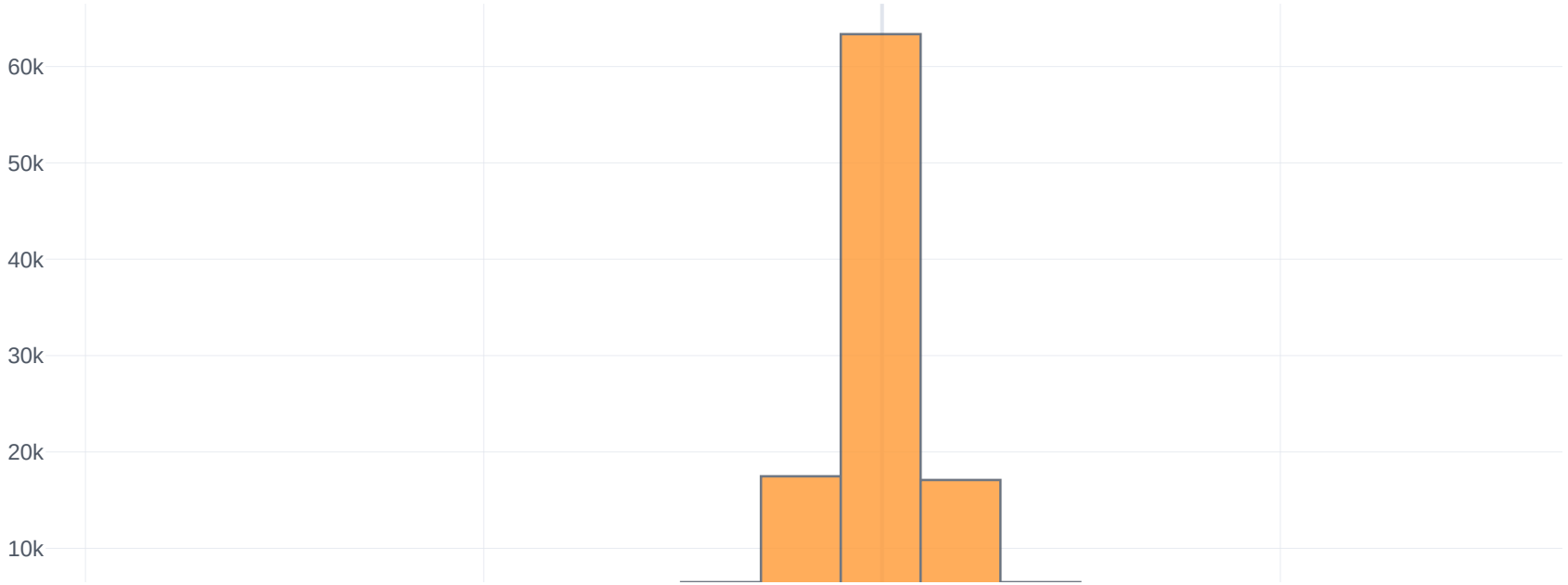



suma_promedio_number_goals's histogram
7000
6000
5000
4000
3000
2000
1000

diferencia_promedio_number_goals's histogram

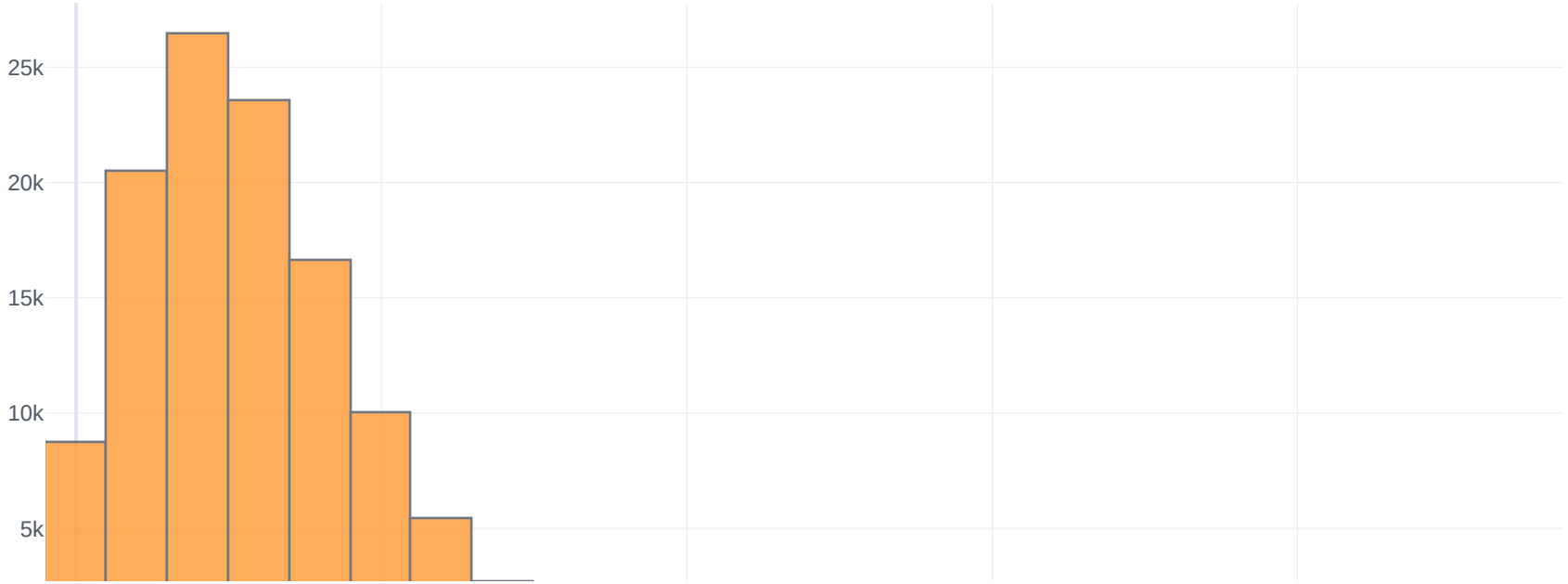


diferencia_sustituciones's histogram

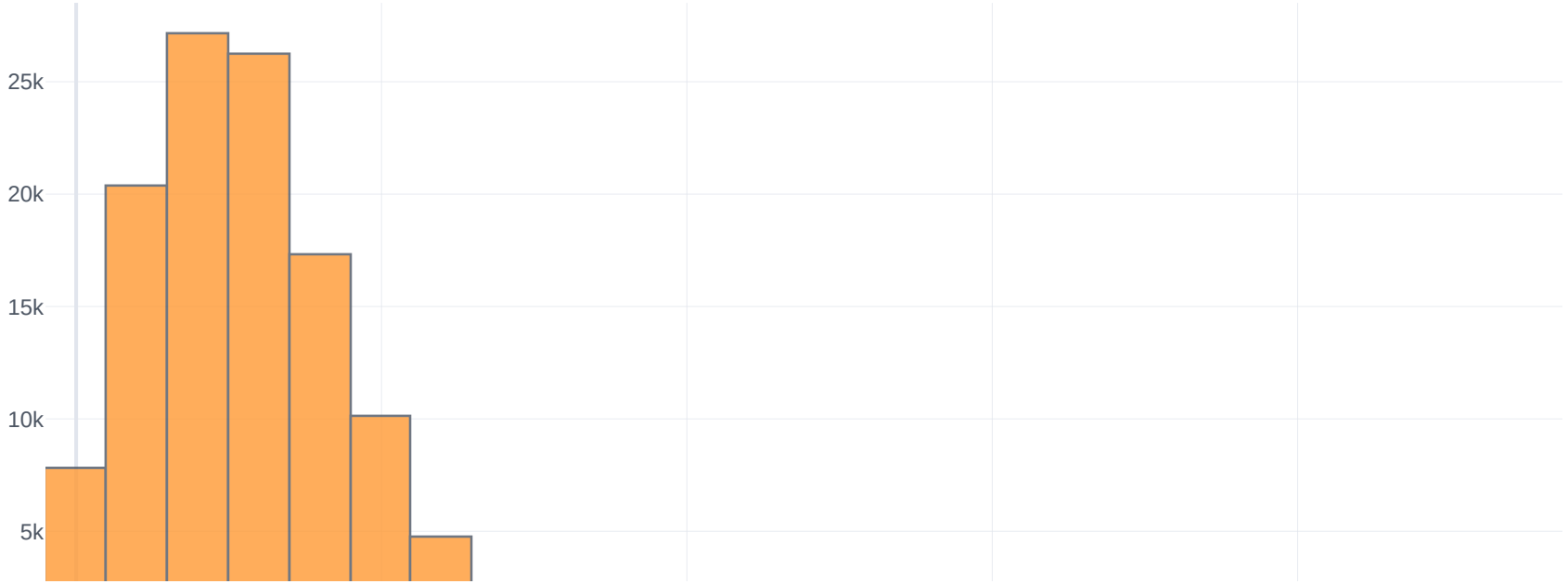


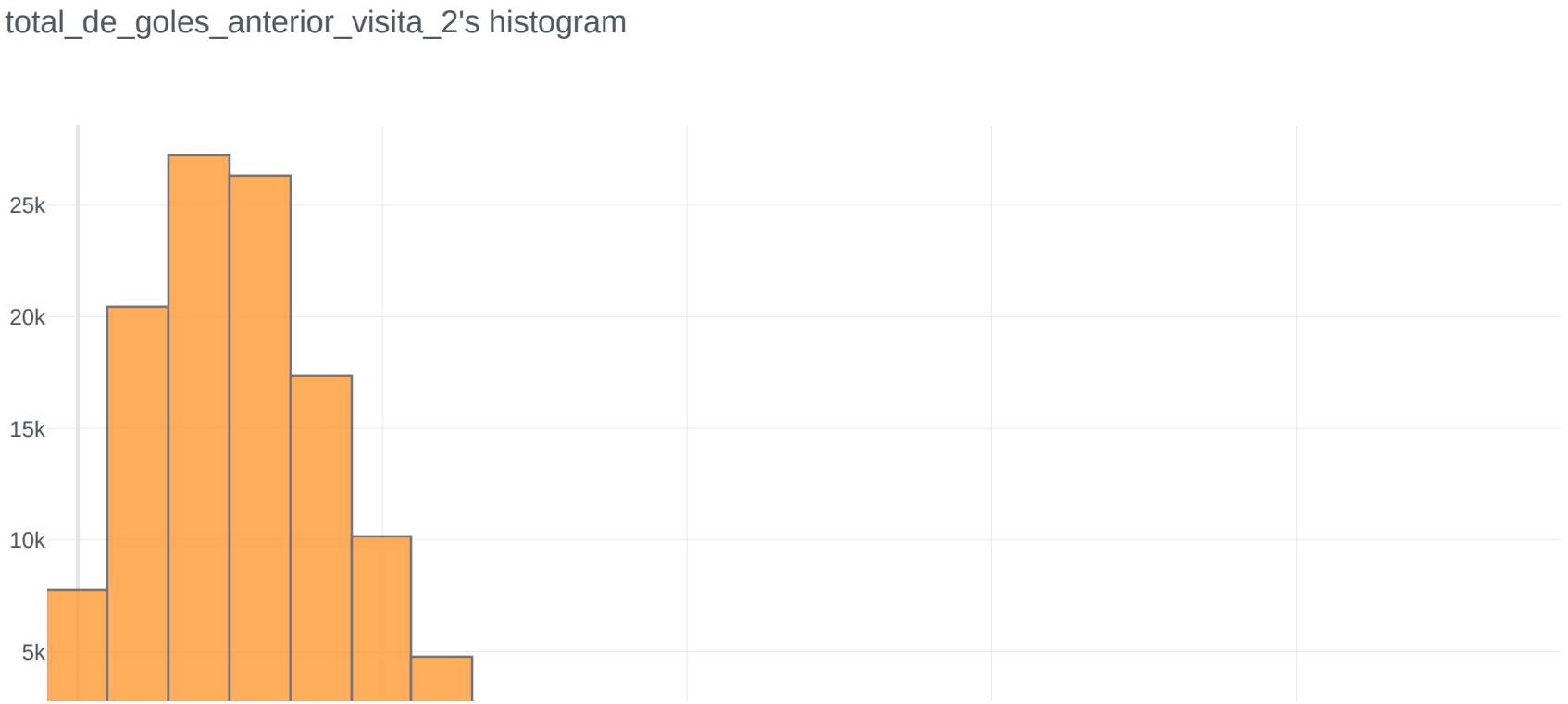
suma_goles_recibidos_2's histogram

suma_goles_anotados_2's histogram

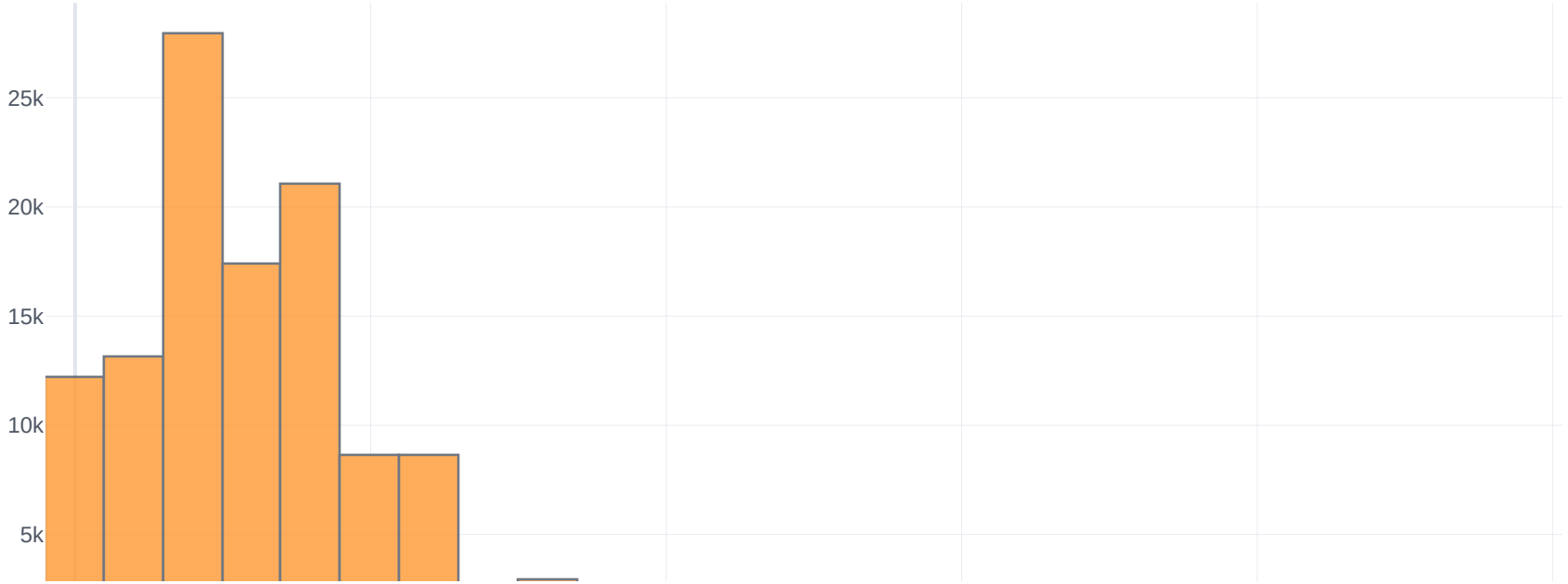


total_de_goles_anterior_casa_2's histogram



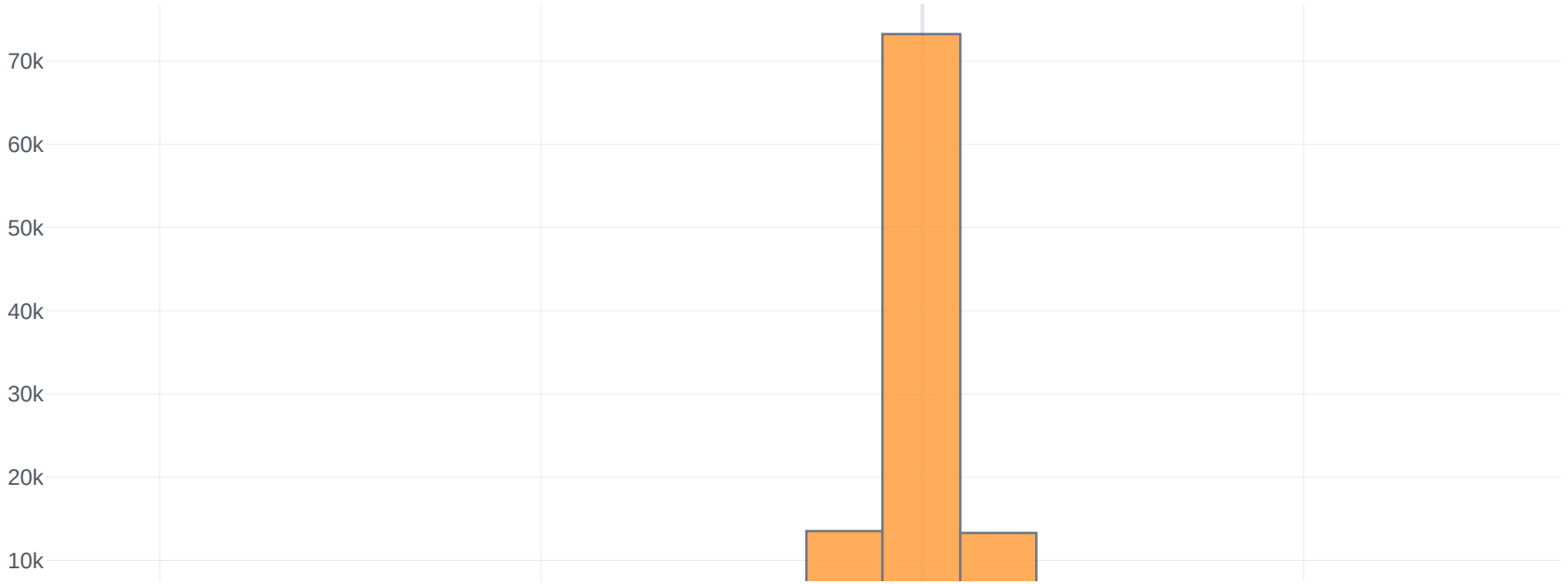

total_de_goles_anterior_visita_2's histogram
25k
20k
15k
10k
5k

anotado_local_recibido_visita_2's histogram



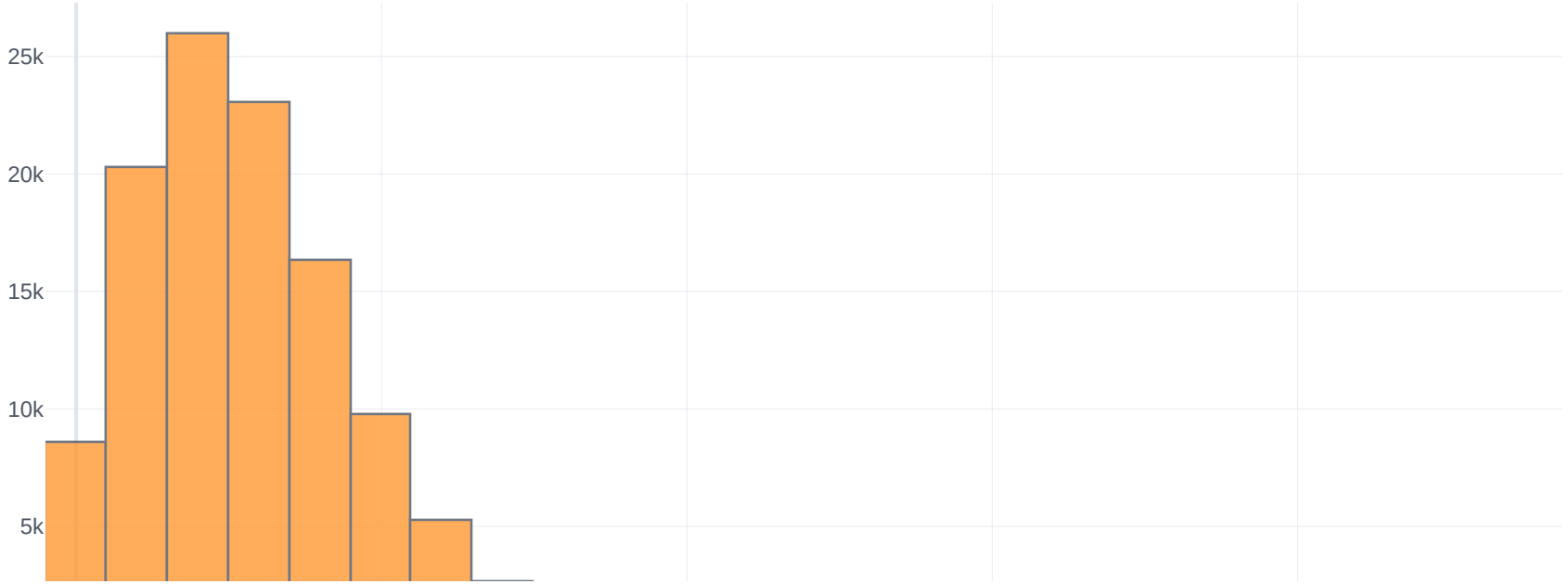
anotado_visita_recibido_local_2's histogram

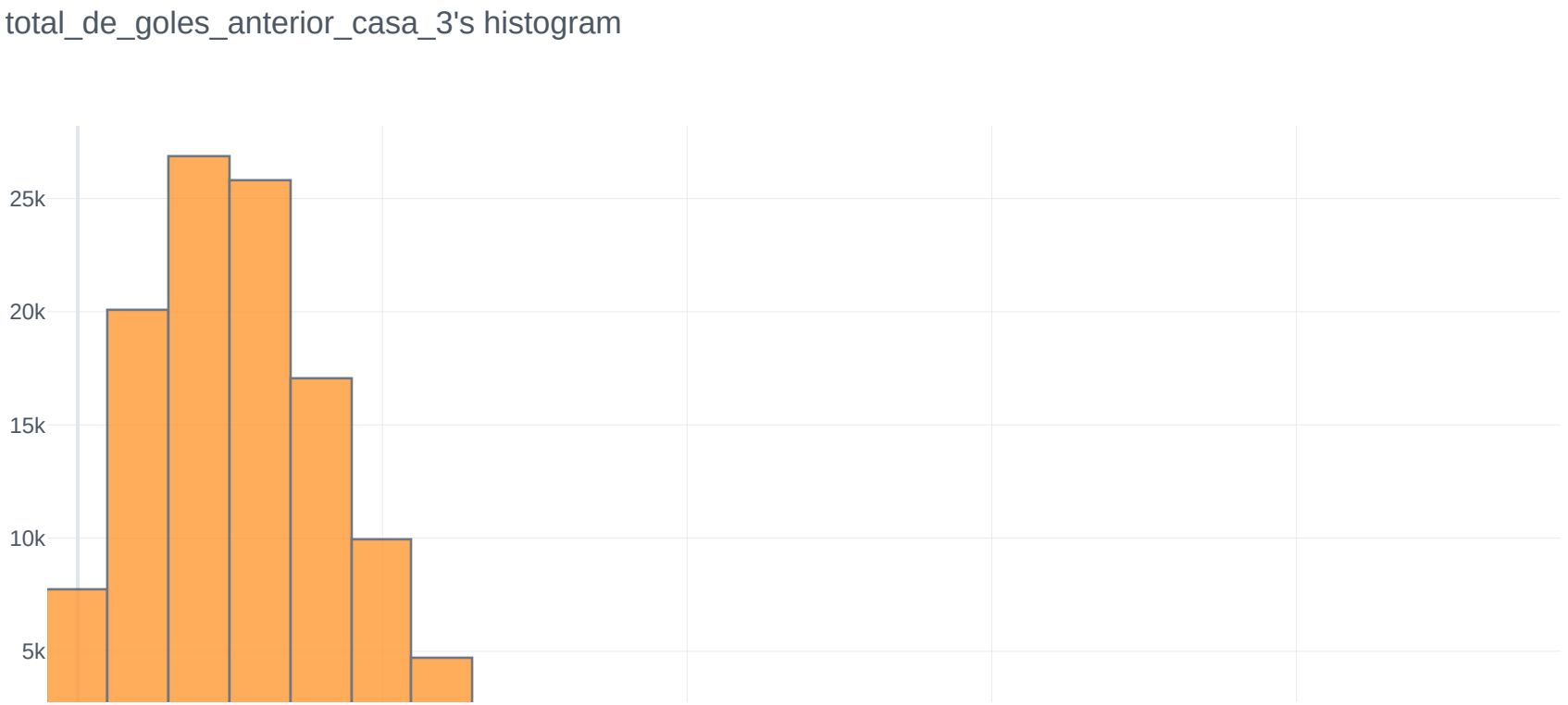
diferencia_sustituciones_2's histogram

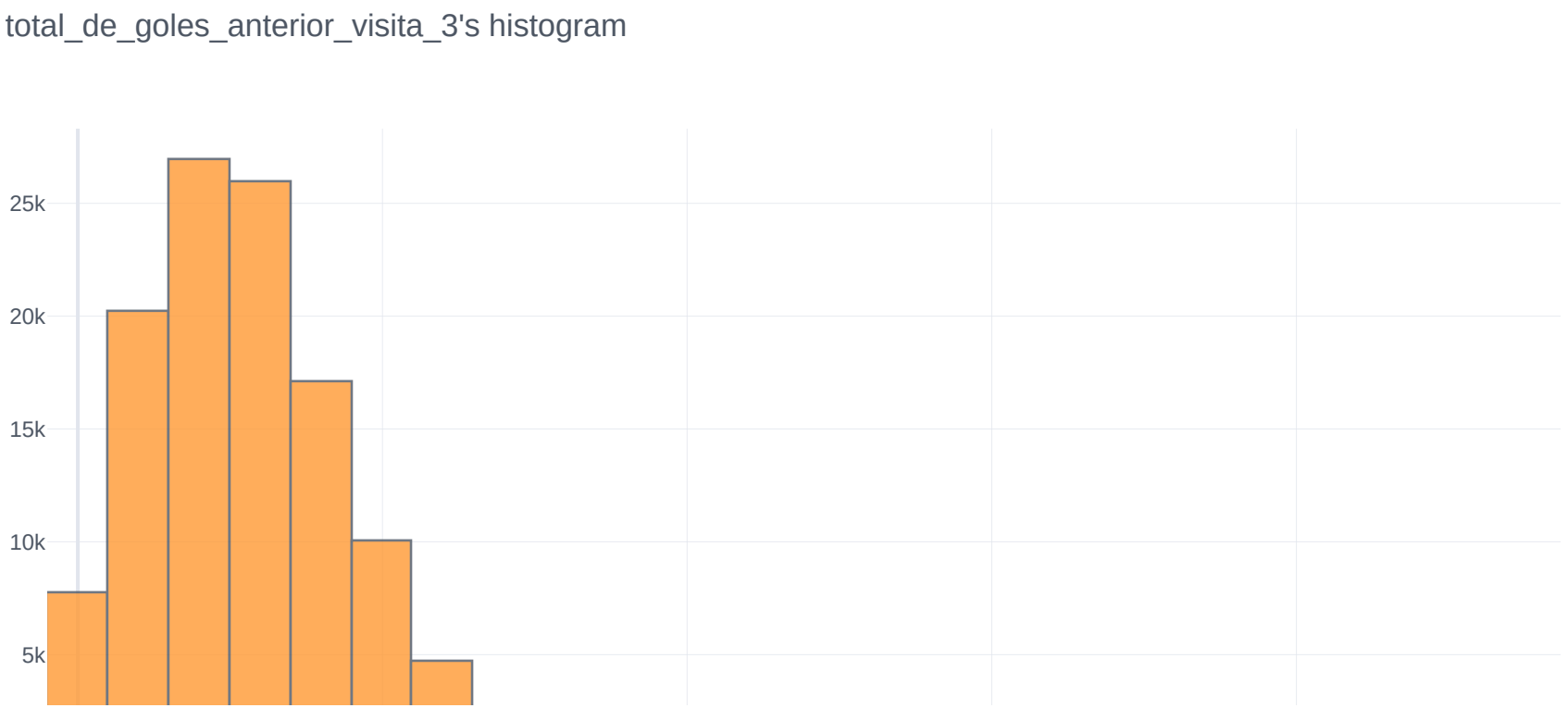


suma_goles_recibidos_3's histogram

suma_goles_anotados_3's histogram

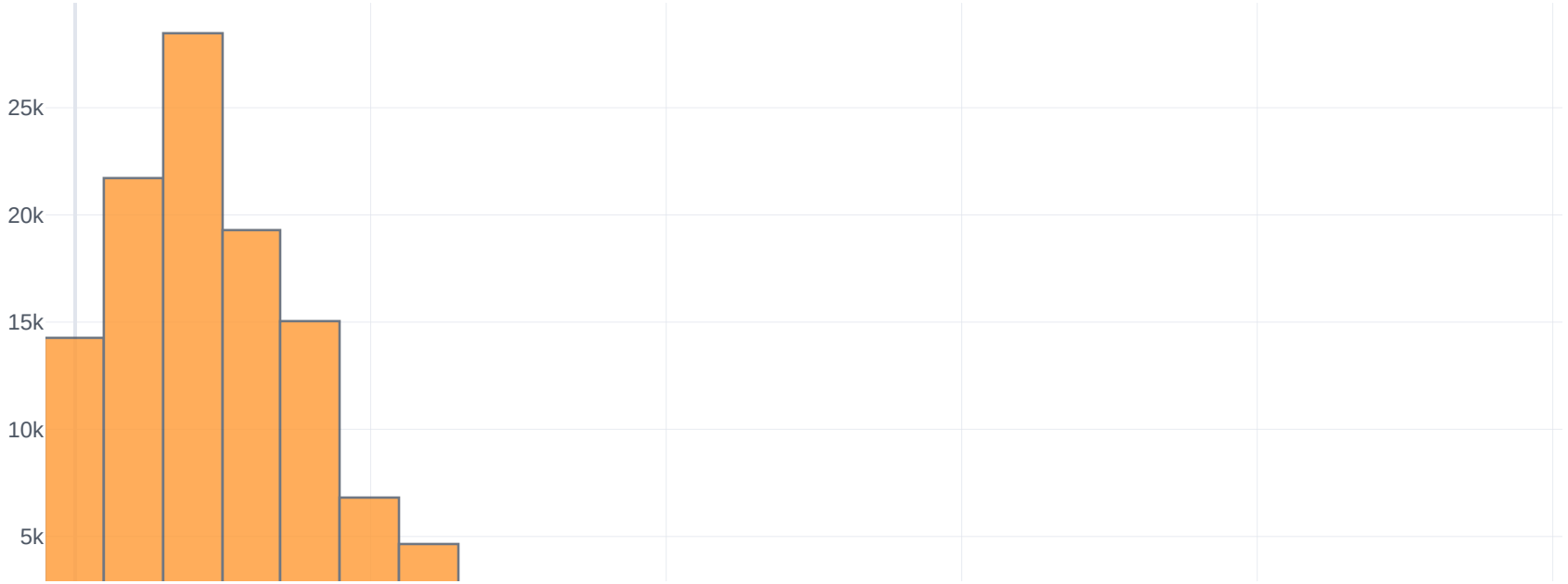


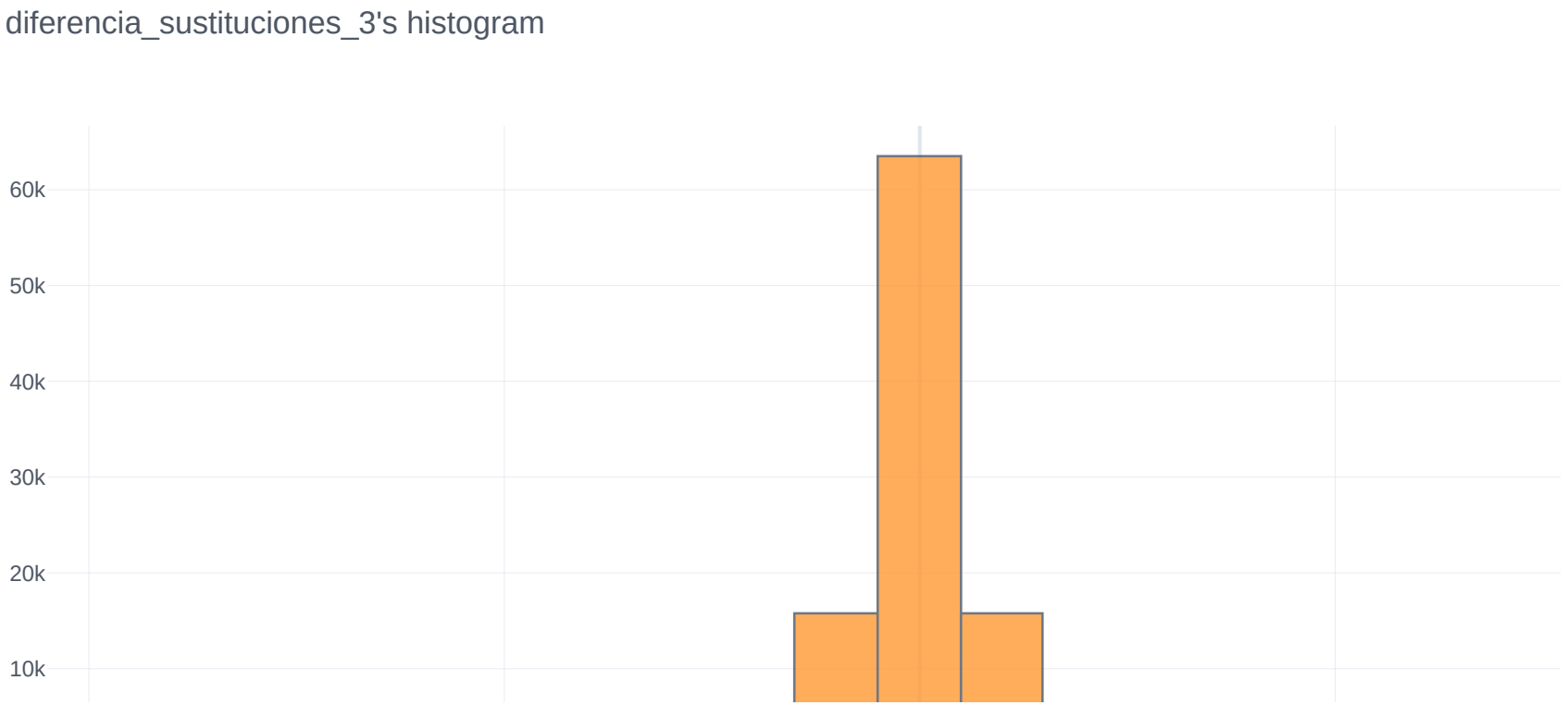

total_de_goles_anterior_casa_3's histogram
25k
20k
15k
10k
5k


total_de_goles_anterior_visita_3's histogram
25k
20k
15k
10k
5k

anotado_local_recibido_visita_3's histogram

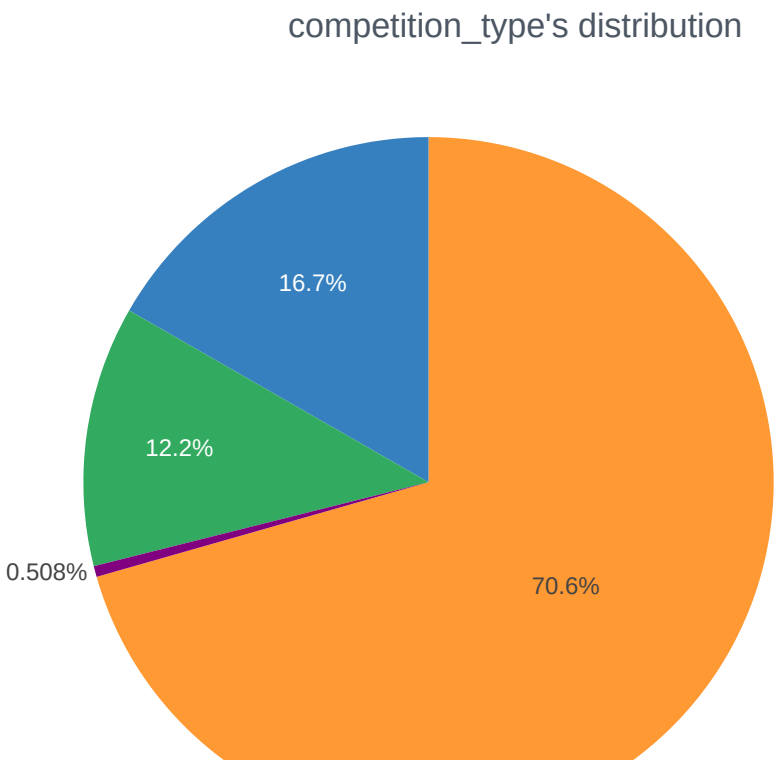
anotado_visita_recibido_local_3's histogram

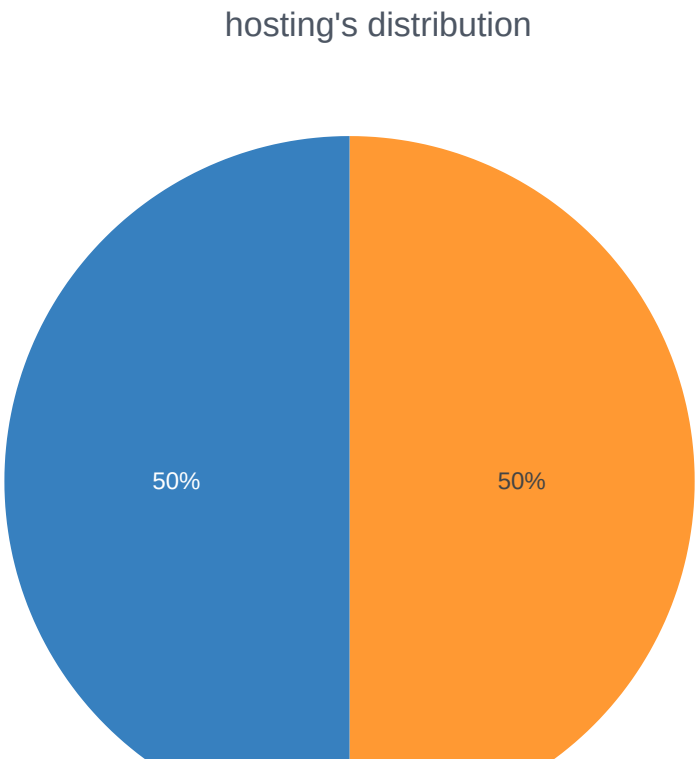


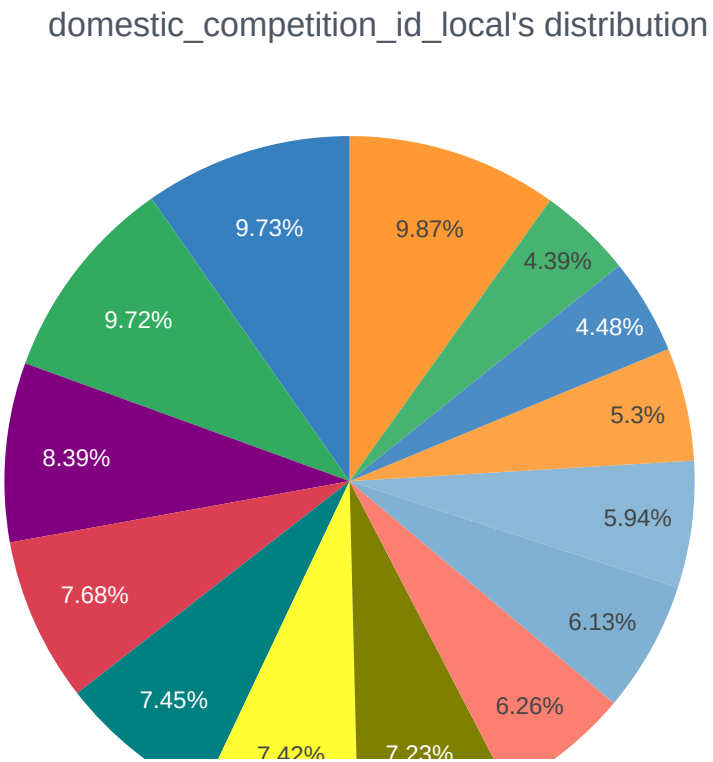


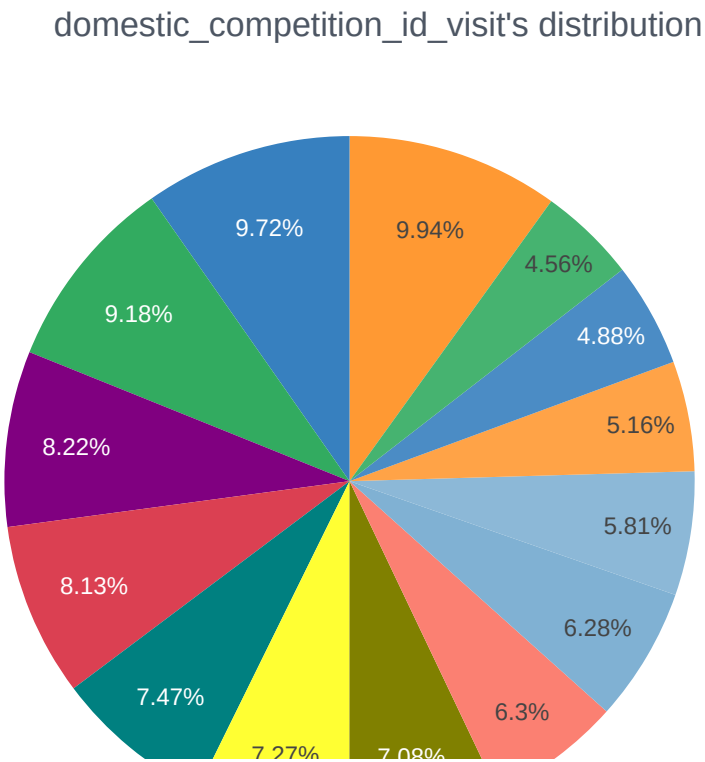
In [32]:
```
# Distrinucion de variables discretas
ls_disc=[]
for i in merge_table.columns:
        if merge_table[i].nunique() < 30:
            if i in ls_cont:
                continue
            else:
                ls_disc.append(i)
```

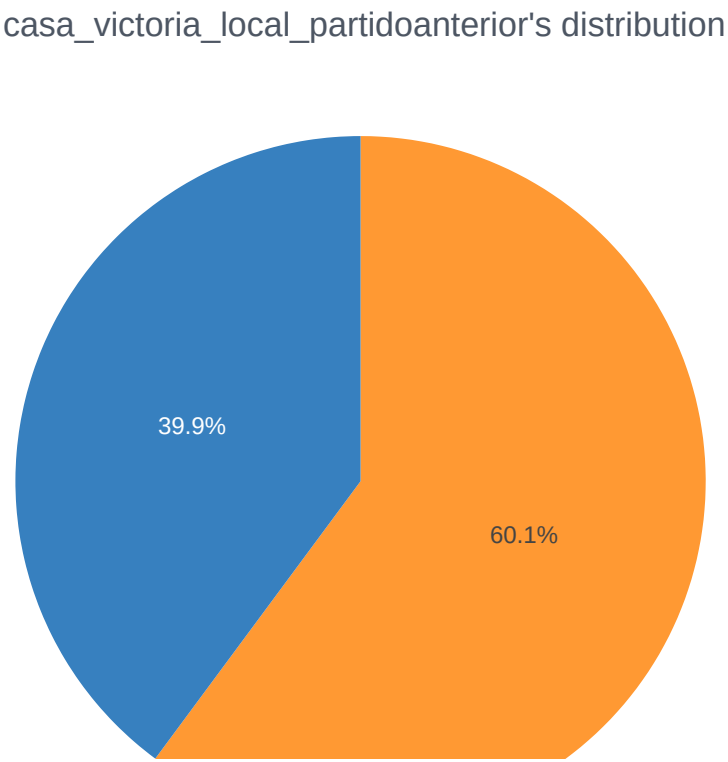
```
for variable in ls_disc:
    merge_table[variable].value_counts().reset_index().iplot(kind="pie",labels=variable, values="count", title=f"{v
```



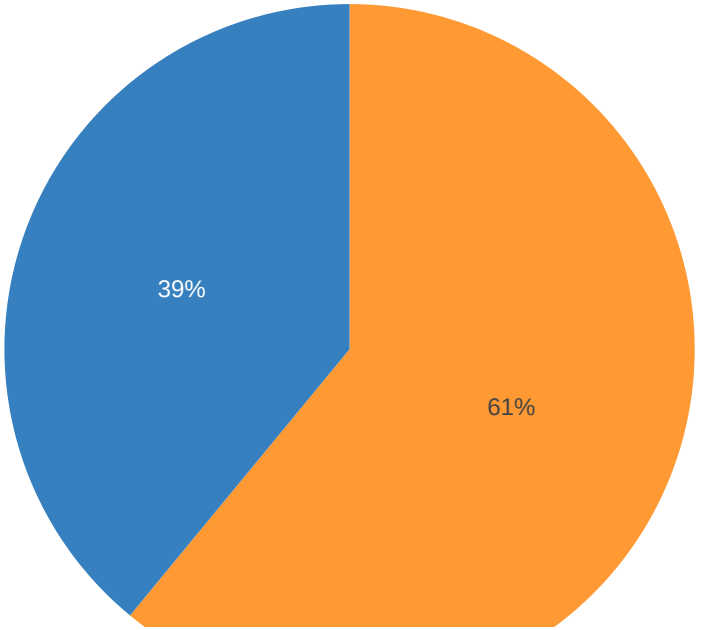

hosting's distribution
50%
50%

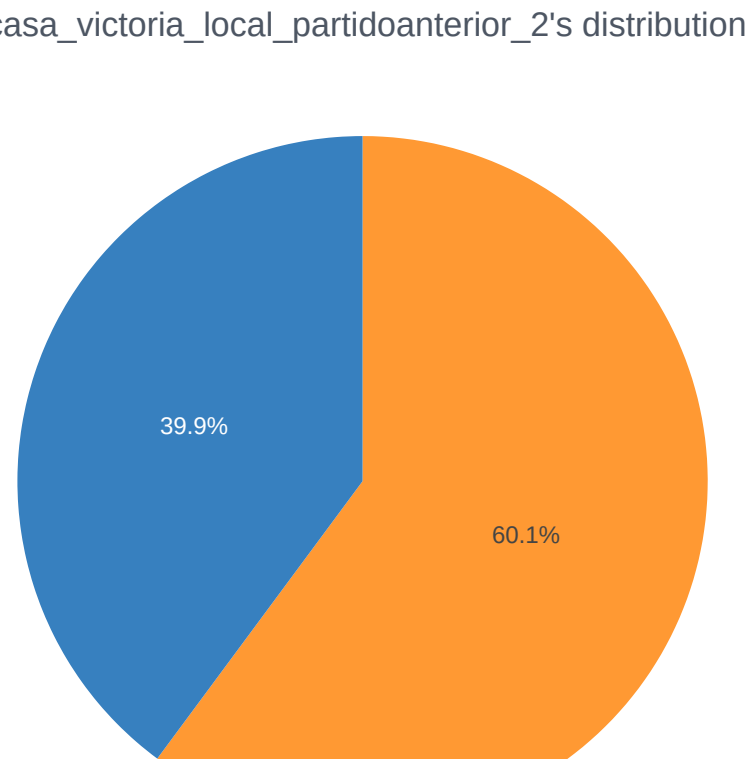

domestic_competition_id_local's distribution
9.73%
9.87%
4.39%
4.48%
9.72%
5.3%
8.39%
5.94%
7.68%
6.13%
7.45%
6.26%
7.42%
7.23%

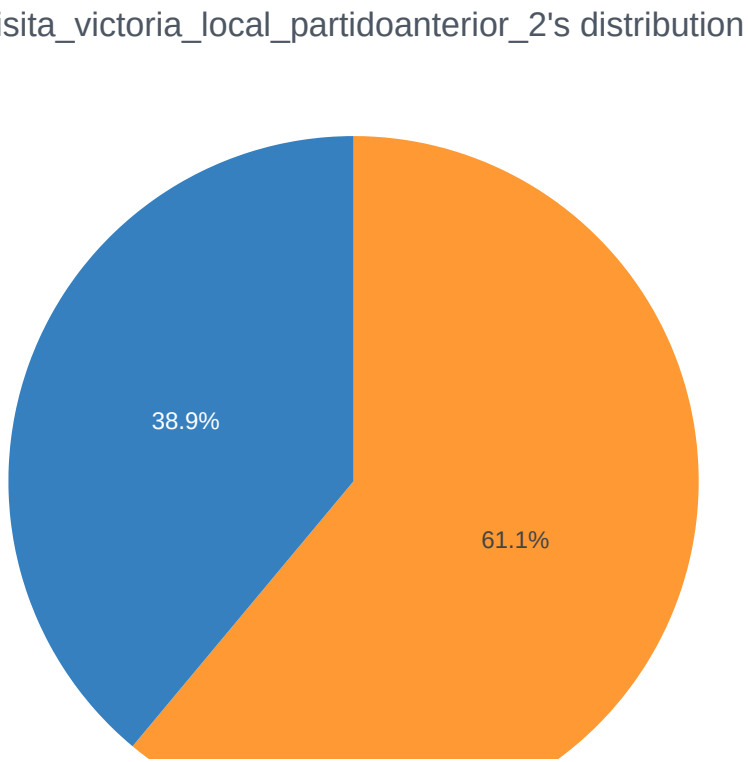

domestic_competition_id_visit's distribution
9.72%
9.94%
4.56%
4.88%
9.18%
5.16%
8.22%
5.81%
8.13%
6.28%
7.47%
6.3%
7.27%
7.08%

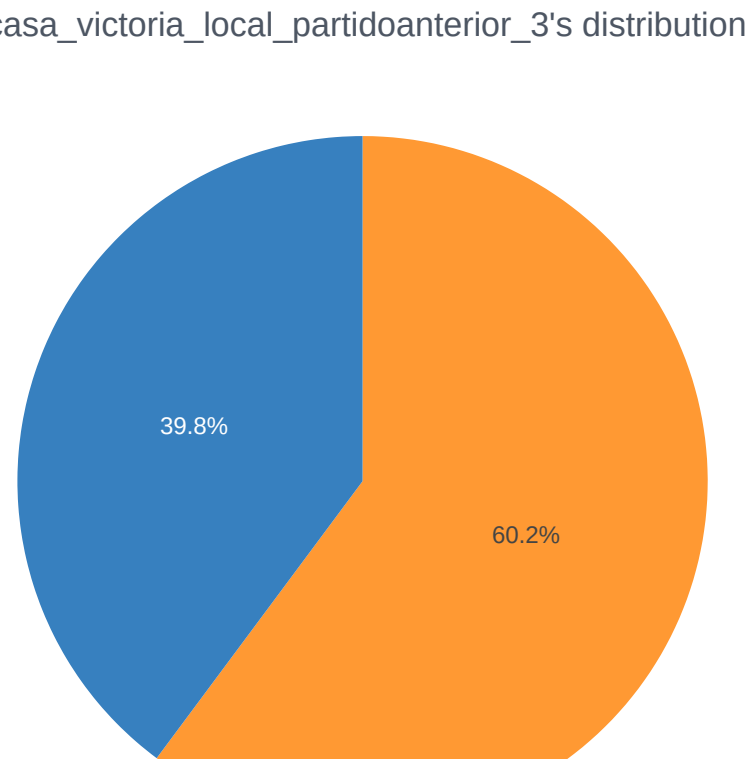

casa_victoria_local_partidoanterior's distribution
39.9%
60.1%

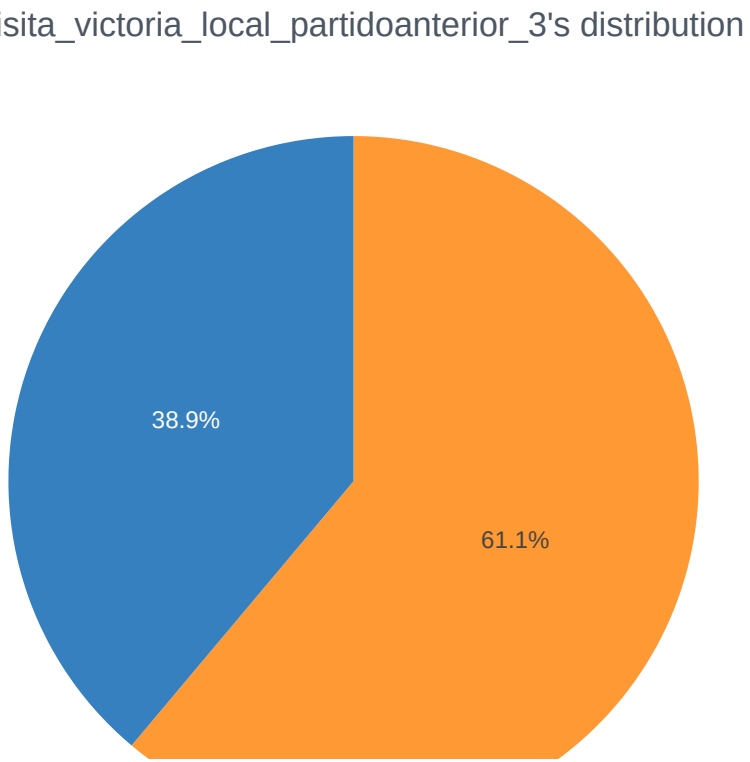
visita_victoria_local_partidoanterior's distribution

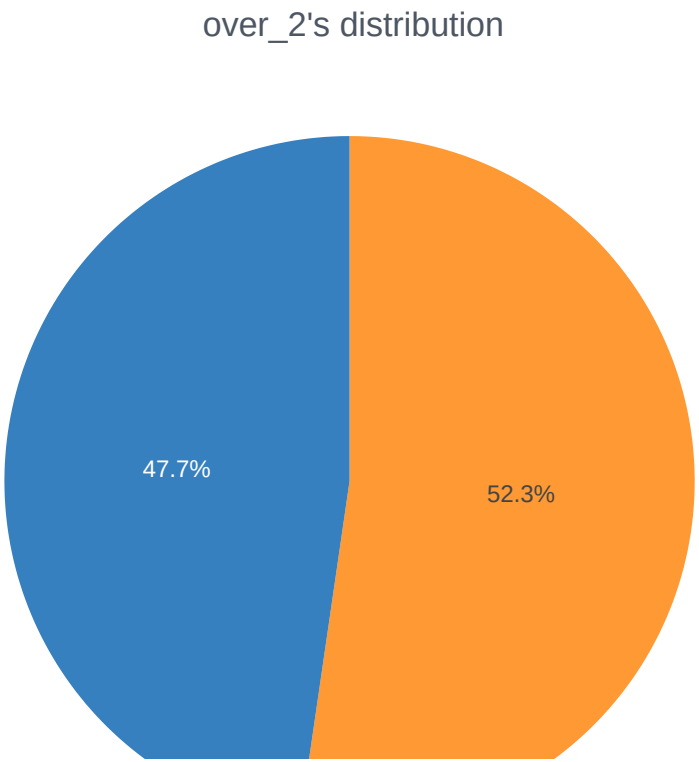


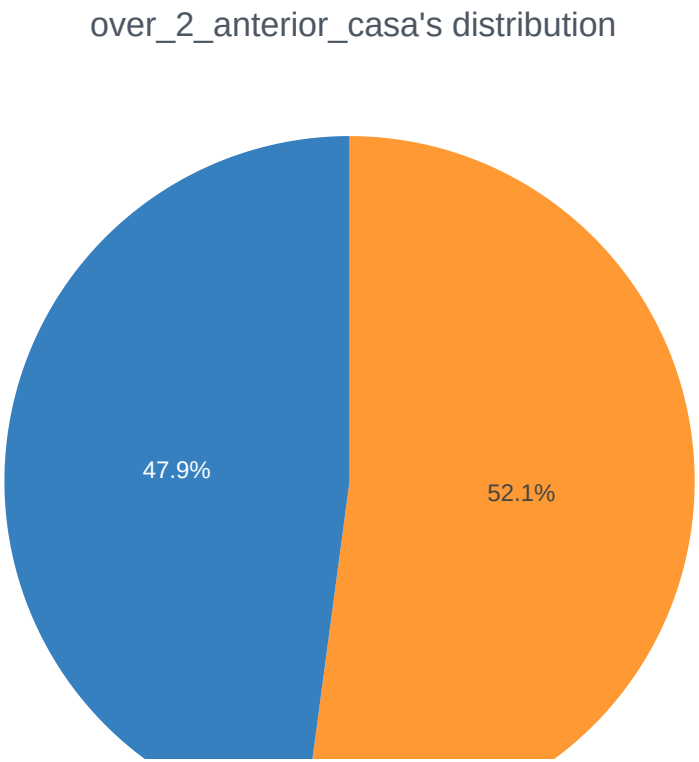

casa_victoria_local_partidoanterior_2's distribution
39.9%
60.1%

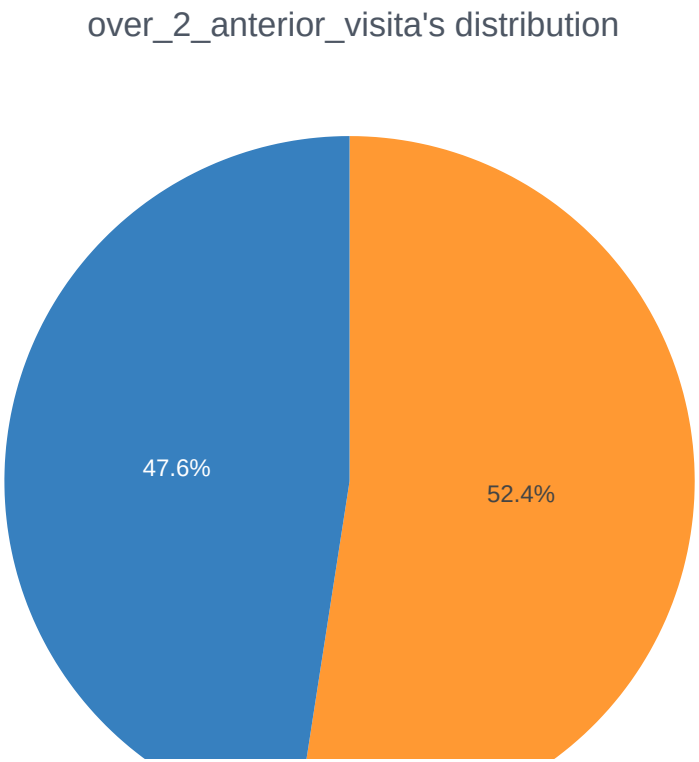

visita_victoria_local_partidoanterior_2's distribution
38.9%
61.1%

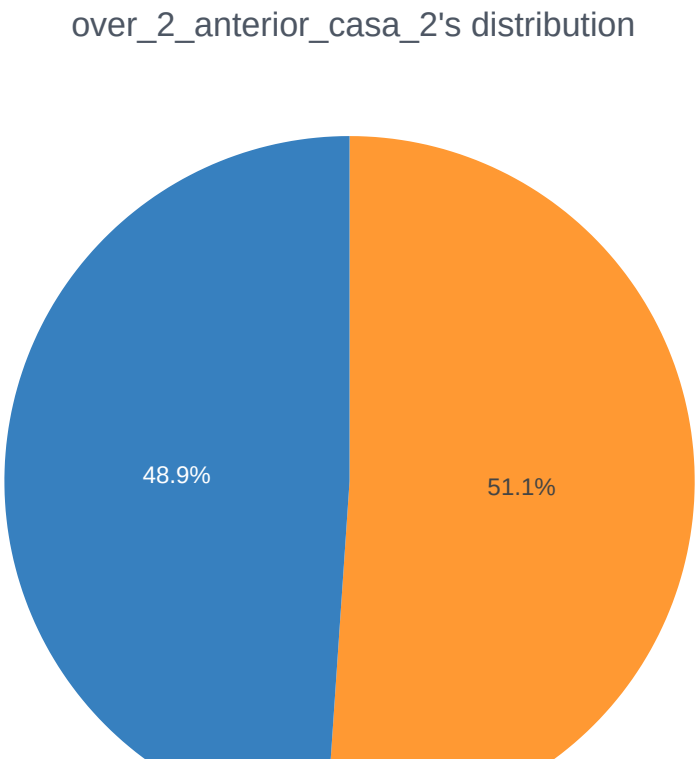

casa_victoria_local_partidoanterior_3's distribution
39.8%
60.2%

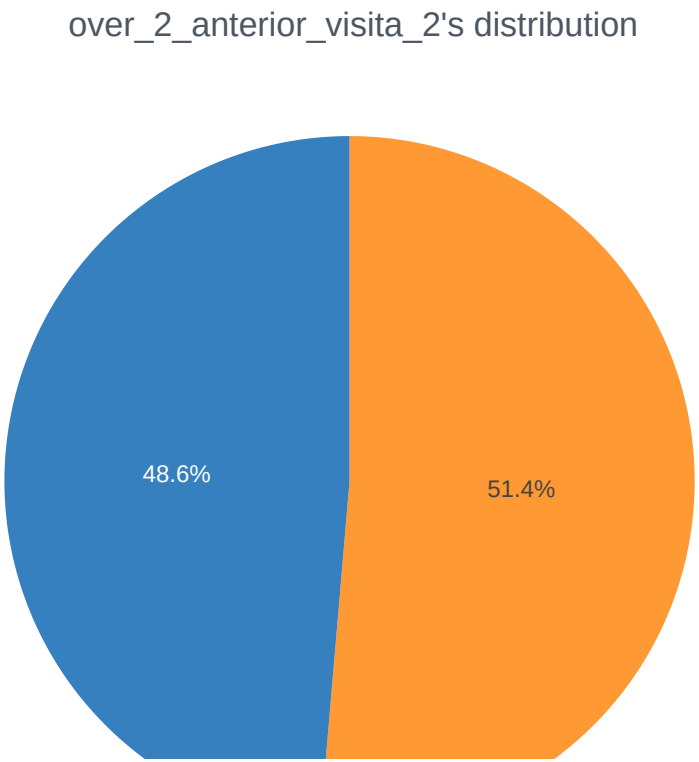

visita_victoria_local_partidoanterior_3's distribution
38.9%
61.1%

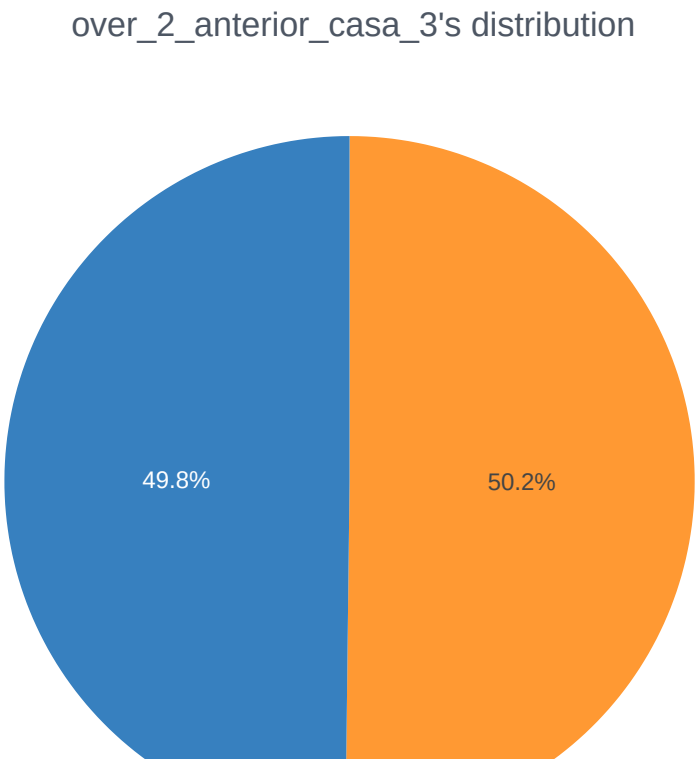

over_2's distribution
47.7%
52.3%

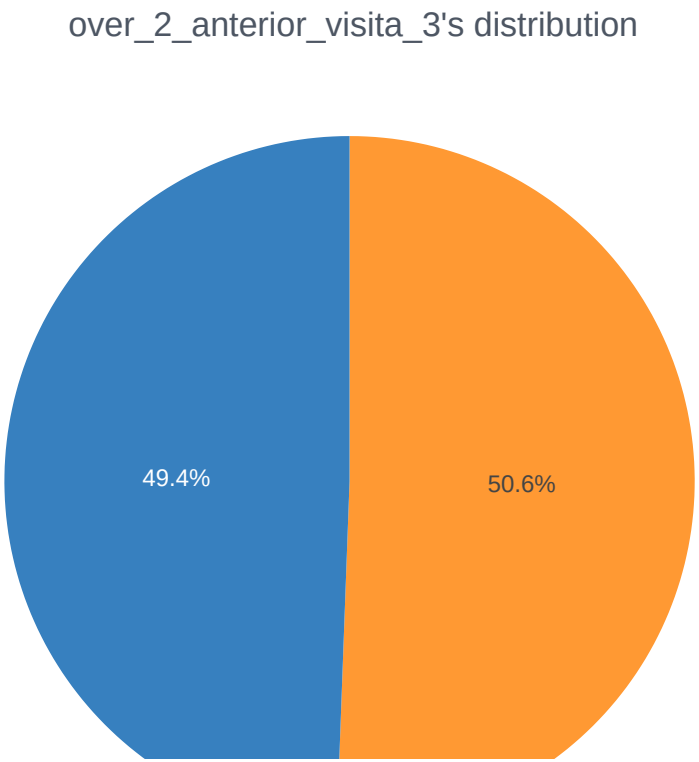

over_2_anterior_casa's distribution
47.9%
52.1%


over_2_anterior_visita's distribution
47.6%
52.4%


over_2_anterior_casa_2's distribution
48.9%
51.1%


over_2_anterior_visita_2's distribution
48.6%
51.4%


over_2_anterior_casa_3's distribution
49.8%
50.2%



In [ ]:
```
#  Se comento ya que el notebook pesaba demasiado, pero se incluye en el reporte
"""# Distribucion de variales en relacion con la variable objetivo
import plotly.graph_objs as go
from plotly.offline import iplot
import plotly.offline as pyo
for variable in ls_cont:
    data = []
```

```
    trace1 = go.Histogram(
        x=merge_table[merge_table['over_2'] == 0][variable],
        name='Categoría 0',
        histnorm='percent',
        opacity=0.5
    )
    data.append(trace1)

    trace2 = go.Histogram(
        x=merge_table[merge_table['over_2'] == 1][variable],
        name='Categoría 1',
        histnorm='percent',
        opacity=0.5
    )
    data.append(trace2)

    layout = go.Layout(
        title='Histograma de ' + variable + ' con respecto a la Categoría ' + 'over_2',
        xaxis=dict(title='Valor'),
        yaxis=dict(title='Frecuencia %'),
        barmode='overlay'
    )

    fig = go.Figure(data=data, layout=layout)
    pyo.iplot(fig)""" #descomentar para eda
```

In [33]:
```
merge_table.info()
```

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 122120 entries, 2012-07-03 to 2023-03-19
Columns: 137 entries, away_club_id to diferencia_sustituciones_3
dtypes: float64(96), int64(16), object(25)
memory usage: 128.6+ MB
```

In [34]:
```
ls_disc
```

Out[34]:
```
['competition_type',
 'hosting',
 'domestic_competition_id_local',
 'domestic_competition_id_visit',
 'casa_victoria_local_partidoanterior',
 'visita_victoria_local_partidoanterior',
 'casa_victoria_local_partidoanterior_2',
 'visita_victoria_local_partidoanterior_2',
 'casa_victoria_local_partidoanterior_3',
 'visita_victoria_local_partidoanterior_3',
 'over_2',
 'over_2_anterior_casa',
 'over_2_anterior_visita',
 'over_2_anterior_casa_2',
 'over_2_anterior_visita_2',
 'over_2_anterior_casa_3',
 'over_2_anterior_visita_3']
```

In [35]:
```
#eliminasmos la variable "hosting"
merge_table=merge_table.drop(["hosting"], axis=1)
```

In [36]:
```
ls_disc.remove("hosting")
```

Creamos dummies para las variables discretas tipo string

In [37]:
```
# dummies de la variable subject
ls_dummies=[]
for i in ls_disc:
    if merge_table[i].dtype == object:
        ls_dummies.append(i)
dumm=pd.get_dummies(merge_table[ls_dummies],dtype=int,drop_first=True) # se usa tipo entero y se elimina una de las
for i in dumm.columns: # se agregan las columnas al df original
    merge_table[i]=dumm[i]
merge_table.info()
```

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 122120 entries, 2012-07-03 to 2023-03-19
Columns: 165 entries, away_club_id to domestic_competition_id_visit_UKR1
dtypes: float64(96), int64(45), object(24)
memory usage: 154.7+ MB
```

In [38]:
```python
# actualizamos la lista de variables discretas
ls_disc=[]
for i in merge_table.columns:
    if i in ls_cont:
        continue
    else:
        if not merge_table[i].dtype == object:
            ls_disc.append(i)
len(ls_disc)
```

Out[38]: 42

Limpieza

In [39]:
```python
# vamos a quitar las variables que no nos interesan de ls_cont
ls_cont.remove('away_club_id')
ls_cont.remove('club_id')
```

In [40]:
```python
#vemos que no tenemos valores unarios para este caso
merge_table[ls_cont+ls_disc].nunique().sort_values().to_frame() #vemos que ya no hay mas vairbales unarias
```

Out[40]:

	0
domestic_competition_id_visit_UKR1	2
domestic_competition_id_local_DK1	2
competition_type_other	2
competition_type_international_cup	2
competition_type_domestic_league	2
...	...
diferencia_promedio_number_goals	570
diferencia_promedio_anotados	585
diferencia_promedio_recibidos	592
foreigners_percentage_difference	2522
attendance	25979

139 rows × 1 columns

In [41]:
```
nulos=merge_table[ls_disc+ls_cont].isnull().mean().sort_values().to_frame().reset_index()
nulos
```

Out[41]:

	index	0
0	domestic_competition_id_visit_GB1	0.000000
1	domestic_competition_id_local_SC1	0.000000
2	domestic_competition_id_local_UKR1	0.000000
3	domestic_competition_id_visit_DK1	0.000000
4	domestic_competition_id_visit_ES1	0.000000
...	...	...
134	foreigners_percentage_local	0.206731
135	age_difference	0.252964
136	foreigners_percentage_difference	0.269080
137	total_market_value_visit	0.818277
138	total_market_value_local	0.822928

139 rows × 2 columns

In [42]:
```
# actualizamos a nuestro df
merge_table=merge_table[ls_cont+ls_disc]
```

In [43]:
```
# remocion de variables con valores ausentes
shape_df=merge_table.shape
ls_miss_var=[]
nulos=nulos[nulos[0]>.15]
for i in nulos["index"]:
        ls_miss_var.append(f"{i}")
print(f"nuestro conjunto de datos tiene {len(ls_miss_var)} variables con mas del 15% de ausentes y son {ls_miss_var
```

```
nuestro conjunto de datos tiene 15 variables con mas del 15% de ausentes y son ['foreigners_number_visit', 'squad_si
ze_visit', 'national_team_players_visit', 'squad_size_local', 'foreigners_number_local', 'stadium_seats_local', 'nat
ional_team_players_local', 'average_age_visit', 'foreigners_percentage_visit', 'average_age_local', 'foreigners_perc
entage_local', 'age_difference', 'foreigners_percentage_difference', 'total_market_value_visit', 'total_market_value
_local']
```

In [44]:
```
#vamos a eliminar todas las columnas con mas del 10% de valores ausentes
ls_position=["home_club_position","away_club_position"]#notamos que estas varibles no tienen nulos, pero el 25 de l
for i in ls_miss_var+ls_position:
    merge_table=merge_table.drop([i],axis=1)
```

In [45]:
```
#ahora solo tenemos 34 columnas pero ninguna tiene mas del 10% de valores ausentes, sin embargo debemos de darle un
merge_table.info()
```

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 122120 entries, 2012-07-03 to 2023-03-19
Columns: 122 entries, season to domestic_competition_id_visit_UKR1
dtypes: float64(81), int64(41)
memory usage: 114.6 MB
```

In [46]:
```
# los demas valores nulos son resultado de que las variables que creamos,son apartir de otros registros anteriores,
#de cada equipo como local y visita tendra valor nulo al no tener valor anterior, por lo que podemos desasernos de
merge_table=merge_table.dropna()
merge_table.info()
```

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 109939 entries, 2012-07-24 to 2023-03-19
Columns: 122 entries, season to domestic_competition_id_visit_UKR1
dtypes: float64(81), int64(41)
memory usage: 103.2 MB
```

In [47]:
```
# vamos a actualizar la lista ls_cont y ls_disc
ls_cont=[]
for i in merge_table.columns:
    if merge_table[i].nunique()>2:
        ls_cont.append(i)
ls_disc=[]
for i in merge_table.columns:
    if not i in ls_cont:
        ls_disc.append(i)
print(ls_cont,len(ls_cont))
print(ls_disc, len(ls_disc))
```

```
['season', 'attendance', 'own_position', 'opponent_position', 'number_goals', 'casa_goles_anteriores', 'casa_goles_a
nteriores_recibidos', 'casa_minuto_primer_gol_anterior', 'casa_minuto_ultimo_gol_anterior', 'casa_number_goals_anter
ior', 'casa_sustituciones_anterior', 'casa_goles_tiempo_añadido_anterior', 'visita_goles_anteriores_recibidos', 'vis
ita_goles_anteriores', 'visita_minuto_primer_gol_anterior', 'visita_minuto_ultimo_gol_anterior', 'visita_number_goal
s_anterior', 'visita_sustituciones_anterior', 'visita_goles_tiempo_añadido_anterior', 'casa_goles_anteriores_2', 'ca
sa_goles_anteriores_recibidos_2', 'casa_minuto_primer_gol_anterior_2', 'casa_minuto_ultimo_gol_anterior_2', 'casa_nu
mber_goals_anterior_2', 'casa_sustituciones_anterior_2', 'casa_goles_tiempo_añadido_anterior_2', 'visita_goles_anter
iores_recibidos_2', 'visita_goles_anteriores_2', 'visita_minuto_primer_gol_anterior_2', 'visita_minuto_ultimo_gol_an
terior_2', 'visita_number_goals_anterior_2', 'visita_sustituciones_anterior_2', 'visita_goles_tiempo_añadido_anterio
r_2', 'casa_goles_anteriores_3', 'casa_goles_anteriores_recibidos_3', 'casa_minuto_primer_gol_anterior_3', 'casa_min
uto_ultimo_gol_anterior_3', 'casa_number_goals_anterior_3', 'casa_sustituciones_anterior_3', 'casa_goles_tiempo_añad
ido_anterior_3', 'visita_goles_anteriores_recibidos_3', 'visita_goles_anteriores_3', 'visita_minuto_primer_gol_anter
ior_3', 'visita_minuto_ultimo_gol_anterior_3', 'visita_number_goals_anterior_3', 'visita_sustituciones_anterior_3',
'visita_goles_tiempo_añadido_anterior_3', 'casa_goles_recibidos_mean', 'casa_goles_anotados_mean', 'visita_goles_rec
ibidos_mean', 'visita_goles_anotados_mean', 'casa_number_goals_mean', 'visita_number_goals_mean', 'suma_goles_recibi
dos', 'suma_goles_anotados', 'total_de_goles_anterior_casa', 'total_de_goles_anterior_visita', 'anotado_local_recibi
do_visita', 'anotado_visita_recibido_local', 'diferencia_promedio_anotados', 'diferencia_promedio_recibidos', 'suma_
promedio_recibidos_anotados', 'suma_promedio_anotados_recibidos', 'suma_promedio_number_goals', 'diferencia_promedio
_number_goals', 'diferencia_sustituciones', 'suma_goles_recibidos_2', 'suma_goles_anotados_2', 'total_de_goles_anter
ior_casa_2', 'total_de_goles_anterior_visita_2', 'anotado_local_recibido_visita_2', 'anotado_visita_recibido_local_
2', 'diferencia_sustituciones_2', 'suma_goles_recibidos_3', 'suma_goles_anotados_3', 'total_de_goles_anterior_casa_
3', 'total_de_goles_anterior_visita_3', 'anotado_local_recibido_visita_3', 'anotado_visita_recibido_local_3', 'difer
encia_sustituciones_3'] 80
['casa_victoria_local_partidoanterior', 'visita_victoria_local_partidoanterior', 'casa_victoria_local_partidoanterio
r_2', 'visita_victoria_local_partidoanterior_2', 'casa_victoria_local_partidoanterior_3', 'visita_victoria_local_par
tidoanterior_3', 'over_2', 'over_2_anterior_casa', 'over_2_anterior_visita', 'over_2_anterior_casa_2', 'over_2_anter
ior_visita_2', 'over_2_anterior_casa_3', 'over_2_anterior_visita_3', 'competition_type_domestic_league', 'competitio
n_type_international_cup', 'competition_type_other', 'domestic_competition_id_local_DK1', 'domestic_competition_id_l
ocal_ES1', 'domestic_competition_id_local_FR1', 'domestic_competition_id_local_GB1', 'domestic_competition_id_local_
GR1', 'domestic_competition_id_local_IT1', 'domestic_competition_id_local_L1', 'domestic_competition_id_local_NL1',
'domestic_competition_id_local_PO1', 'domestic_competition_id_local_RU1', 'domestic_competition_id_local_SC1', 'dome
stic_competition_id_local_TR1', 'domestic_competition_id_local_UKR1', 'domestic_competition_id_visit_DK1', 'domestic
_competition_id_visit_ES1', 'domestic_competition_id_visit_FR1', 'domestic_competition_id_visit_GB1', 'domestic_comp
etition_id_visit_GR1', 'domestic_competition_id_visit_IT1', 'domestic_competition_id_visit_L1', 'domestic_competitio
n_id_visit_NL1', 'domestic_competition_id_visit_PO1', 'domestic_competition_id_visit_RU1', 'domestic_competition_id_
visit_SC1', 'domestic_competition_id_visit_TR1', 'domestic_competition_id_visit_UKR1'] 42
```

In [48]:
```
#vamos a remover las variables altamente correlacionadas
corr_matrix = merge_table[ls_cont].corr()
ls_checked = []
ls_correlated = []
```

```
for col in corr_matrix.columns:
    ls_checked.append(col)
    ls_correlated += corr_matrix[(corr_matrix[col] >= .90) & (~corr_matrix.index.isin(ls_checked))].index.tolist()
ls_correlated = list(set(ls_correlated))
ls_correlated = [variable for variable in ls_correlated if variable not in ls_disc ]
print(f"las variables con correlacion de 1 con respecto a otr son{ls_correlated}")
if len(ls_correlated) >1:
    merge_table.drop(ls_correlated,axis=1)
```

```
las variables con correlacion de 1 con respecto a otr son[]
```

deteccion de valores extremos

In [49]:
```
#definimos funcion para detectar outliers

def detect_outlier(serie, method):
    if method == "iqr":
        q1 = serie.quantile(.25)
        q3 = serie.quantile(.75)
        iqr = q3-q1
        upper_fence = q3 + 1.5*iqr
        lower_fence = q1 - 1.5*iqr
    elif method == "z-score":
        mean = serie.mean()
        std = serie.std()
        upper_fence = mean + 3*std
        lower_fence = mean - 3*std
    else:
        upper_fence = serie.quantile(.99)
        lower_fence = serie.quantile(.01)
    return ~serie.between(lower_fence, upper_fence, inclusive="both")
```

In [50]:
```
Xp = pd.concat(map(lambda column: detect_outlier(merge_table[column], "other").rename(f"{column}_ol"), ls_cont), ax
```

In [51]:
```
shape_old = merge_table.shape
shape_new=Xp[Xp.mean(axis=1)<0.3].shape #renglones que tengan menos del 30% de variables detectadas como outlier
shape_new[0] / shape_old[0] # proporcion de tamaño del nuevo data frame, sin valores extremos
```

Out[51]: 0.9998635607018438

In [52]:
```
#podemos ver que con el metodo que menos perdemos registros, es con el z_core, para este caso nos quedaremos con
#los cuantiles
merge_table =merge_table[Xp.mean(axis=1)<0.3].reset_index(drop=True)
merge_table
```

Out[52]:

	season	attendance	own_position	opponent_position	number_goals	casa_goles_anteriores	casa_goles_anteriores_recibidos	c
0	2011	10000	-1	-1	4	1.0	4.0	
1	2011	6600	-1	-1	7	1.0	0.0	
2	2011	1480	-1	-1	6	7.0	0.0	
3	2011	4653	-1	-1	3	1.0	3.0	
4	2011	3166	-1	-1	1	3.0	2.0	
...	...	...	...	...	...	...	...	
109919	2022	0	2	10	4	1.0	0.0	
109920	2022	3305	12	16	2	0.0	1.0	
109921	2022	10300	1	6	2	1.0	1.0	
109922	2022	9012	16	18	4	1.0	2.0	
109923	2022	9012	18	16	4	2.0	0.0	

109924 rows × 122 columns

In [53]:
```
# variables con infinitos
#Validamos ahora para inf variables
inf_values = np.isinf(merge_table[ls_cont]).sum().items()
inf_variables = [var for var, n_inf in inf_values if n_inf > 0]
inf_variables # no hay variables con infinitos
if len(inf_variables)>1:
    merge_table.drop(inf_variables,axis=1)
```

Separacion de conjuntos

In [54]:
```
x=merge_table.copy().drop(['number_goals','over_2'],axis=1)
y=merge_table["over_2"]
```

In [55]:
```
# separacion en conjuntos de train y test
X_train, X_test, y_train, y_test = train_test_split(x, y, test_size = 1/3, random_state = 0)
```

Escalamiento

In [56]:
```
scaler=MinMaxScaler()
scaler.fit(X_train)
X_scaled=pd.DataFrame(scaler.transform(X_train), columns=X_train.columns)
X_test_scaled=pd.DataFrame(scaler.transform(X_test), columns=X_train.columns)
```

In [57]:
```
X_scaled
```

Out[57]:

	season	attendance	own_position	opponent_position	casa_goles_anteriores	casa_goles_anteriores_recibidos	casa_minuto_pri
0	0.636364	0.241560	0.863636	0.500000	0.000000	0.1250	
1	0.363636	0.059645	0.454545	0.636364	0.000000	0.0000	
2	0.454545	0.018369	0.227273	0.636364	0.266667	0.0000	
3	0.727273	0.025012	0.681818	0.318182	0.200000	0.0625	
4	0.727273	0.104435	0.727273	0.363636	0.200000	0.0000	
...	...	...	...	...	...	...	
73277	0.272727	0.404000	0.636364	0.772727	0.133333	0.1250	
73278	0.454545	0.232834	0.000000	0.000000	0.066667	0.0625	
73279	0.454545	0.045293	0.636364	0.772727	0.066667	0.3125	
73280	0.454545	0.092900	0.318182	0.090909	0.066667	0.1250	
73281	0.636364	0.028434	0.727273	0.136364	0.133333	0.0625	

73282 rows × 120 columns

In [ ]:

```
# para guargar en la computadora
"""X_scaled.to_csv('xs.csv', index=False)
X_test_scaled.to_csv('xs_test.csv', index=False)
y_train.to_csv('y_train.csv', index=False)
y_test.to_csv('y_test.csv', index=False)"""
```