Data Quality Report - Initial Findings

Descriptive Statistics for Continuous Feature

	count	mean	std	min	25%	50%	75%	max
customer	1000.0	1.049027e+06	28337.411453	1.000004e+06	1.025108e+06	1.049217e+06	1.073305e+06	1.099614e+06
age	1000.0	3.063400e+01	22.551599	0.000000e+00	0.000000e+00	3.400000e+01	4.800000e+01	8.600000e+01
numHandsets	1000.0	1.769000e+00	1.365102	1.000000e+00	1.000000e+00	1.000000e+00	2.000000e+00	1.300000e+01
handsetAge	1000.0	4.011510e+02	259.406661	-2.000000e+00	2.180000e+02	3.445000e+02	5.420000e+02	1.812000e+03
currentHandsetPrice	1000.0	3.482581e+01	56.561167	0.000000e+00	0.000000e+00	0.000000e+00	3.249000e+01	3.999900e+02
avgBill	1000.0	5.734780e+01	44.519769	0.000000e+00	3.324750e+01	4.915500e+01	6.917500e+01	4.696700e+02
avgMins	1000.0	5.013036e+02	530.272984	0.000000e+00	1.475000e+02	3.470000e+02	6.638750e+02	4.598750e+03
avgrecurringCharge	1000.0	4.632820e+01	24.246087	0.000000e+00	3.000000e+01	4.499000e+01	5.999000e+01	2.999900e+02
avgOverBundleMins	1000.0	3.888876e+01	94.715038	0.000000e+00	0.000000e+00	4.000000e+00	4.225000e+01	1.389000e+03
avgRoamCalls	1000.0	9.162200e-01	4.940806	0.000000e+00	0.000000e+00	0.000000e+00	2.750000e-01	1.310400e+02
callMinutesChangePct	1000.0	-5.377716e-01	5.197712	-4.465500e+01	-1.856250e+00	-1.750000e-01	9.900000e-01	2.805000e+01
billAmountChangePct	1000.0	-1.048300e-02	0.927263	-5.763000e+00	-1.363000e-01	-5.000000e-03	3.690000e-02	1.791140e+01
avgReceivedMins	1000.0	1.091565e+02	163.394282	0.000000e+00	5.787500e+00	4.970500e+01	1.507550e+02	1.549930e+03
avgOutCalls	1000.0	2.538799e+01	36.119986	0.000000e+00	2.330000e+00	1.283500e+01	3.600000e+01	3.320000e+02
avgInCalls	1000.0	8.439290e+00	17.036525	0.000000e+00	0.000000e+00	2.000000e+00	9.415000e+00	2.330000e+02
peakOffPeakRatio	1000.0	2.137873e+00	3.465256	0.000000e+00	7.226356e-01	1.380353e+00	2.446368e+00	7.477477e+01
peakOffPeakRatioChangePct	1000.0	2.730705e-01	9.620523	-2.815402e+01	-6.099596e+00	5.604797e-02	6.636753e+00	3.777974e+01
avgDroppedCalls	1000.0	9.699970e+00	14.972256	0.000000e+00	1.330000e+00	5.000000e+00	1.141500e+01	1.593300e+02
lifeTime	1000.0	1.874700e+01	9.499433	6.000000e+00	1.100000e+01	1.700000e+01	2.400000e+01	6.000000e+01
lastMonthCustomerCareCalls	1000.0	1.641900e+00	4.167139	0.000000e+00	0.000000e+00	0.000000e+00	1.330000e+00	4.200000e+01
numRetentionCalls	1000.0	5.100000e-02	0.233353	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	2.000000e+00
numRetentionOffersAccepted	1000.0	2.400000e-02	0.159529	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	2.000000e+00
newFrequentNumbers	1000.0	2.040000e-01	0.637802	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	3.000000e+00

We have a full count of all the continuous features.

The age feature seems to have a lot of zero values, as high as 25% of data. Max value is 86 years old, with mean of about 30. Thus it appears the data is covering a wide range of users in terms of age.

The number of handsets seems like a reasonable range of values with 75% of users having had 2 handsets in past 3 years. This is typical of most current mobile phone plans which last 12 to 24 months. The max value is 13 however, which seems like an outlier.

Handset age is presumed to be measured in days rather than months unlike most of the other continuous features which are measured in months. This assumption is justified by dividing the number of days by 365 days. This results in age that's about a year for 50% of people and 1.5 years for 75% of people. As mobile networks often give new phones at the end of one to two year plan, it is not surprising that the phones have an age in range of one to two years. The minimum value is unusual though in that it is negative.

Current handset price has an odd price of zero for 50% of it the customers. This could be due to mobile network offering free phones with their contracts.
The average bill, average minutes, average recurring charge, the average over bundle minutes, average roam calls, average received minutes, average out calls, average in calls, peak/off peak ratio and average dropped calls all have minimum value of zero. This is call data however. The customer could be using the phone to connect to Wi-Fi for e.g. and communicate for free via the Internet. Thus while odd at a glance, I don't believe there is anything unusual about the minimum values of those features.

The number of retention calls and number of retention offers accepted have a lot of zeros with 75% of data being zero. While it's far too early to draw any conclusions now, this has caught my interest as it seems that there may not be enough of retention calls and offers being made to prevent churn.

Descriptive Statistics for Categorical Features

	count	unique	top	freq
occupation	267	7	professional	179
regionType	531	6	suburban	306
marriageStatus	1000	3	unknown	402
children	1000	2	False	765
income	1000	10	0	268
smartPhone	1000	2	True	891
creditRating	1000	7	B	389
homeOwner	1000	2	False	652
creditCard	1000	6	true	655
churn	1000	2	True	532

The occupation count is only 267/1000. Also, 179/267 are "professional", which is arguably not very detailed and doesn't actually tell us a lot.

Similarly, the region type is 531/1000 with 306/531 being "suburban". This doesn't seem to tell us anything useful.

Marriage status is unknown for significant count of 402/1000.

It appears that most of the customers don't have children, as many as 765/1000.

Income is in category 0 for 268/1000 customers. It is unclear if 0 represents high-income or low-income group.

About 90% of the customers appear to have a smartphone, this is as expected in modern day.

About 65% of customers are not home owners and about 65% of customers have a credit card.

The churn overall is true for about 50% of the sample.

Histograms for Continuous Features [plots attached at end of file]

The age appears to be normally distributed with the centre at about 40 years. There is an outlier present where there is a large number of ages at zero.

The average bill, average dropped calls, average in calls, average minutes, average out calls, average over bundle minutes, average received minutes, last month customer care calls, number of handsets and peak/off peak ratio all appear to be exponentially decreasing.

Bill amount change percent and peak/off peak ratio change percent is normally distributed with centre at about 0.

Life time appears very linear.

Average recurring charge and handset age both appear normally distributed and skewed right.

Box Plots for Continuous Features [plots attached at end of file]

Many of the plots have many outliers. Most of the outliers are larger than the max cut off point. Some outliers are pretty extreme, being multiple standard deviations away from the mean. Such examples are number of handsets, handset age, current handset price, average bill, average minutes, average recurring charge, average over bundle minutes, average roam calls, call minute change percent, bill amount change percent, average received minutes, average out calls, average in calls, peak/off peak ratio, average dropped calls and last month customer care calls.

A lot of the outliers do make sense however. As this is mobile data, you can expect for e.g. that if the average minutes have outliers, then those people will pay higher bill and thus the average bill boxplot reflects this with its own high outliers.

Bar Plots for Categorical Features [plots attached at end of file]

Occupation plot is dominated by "professional" category, with other categories roughly equally close to each other. Unfortunately, this doesn't tell us much as the "professional" category is itself too broad, encompassing a wide range of different jobs and backgrounds.

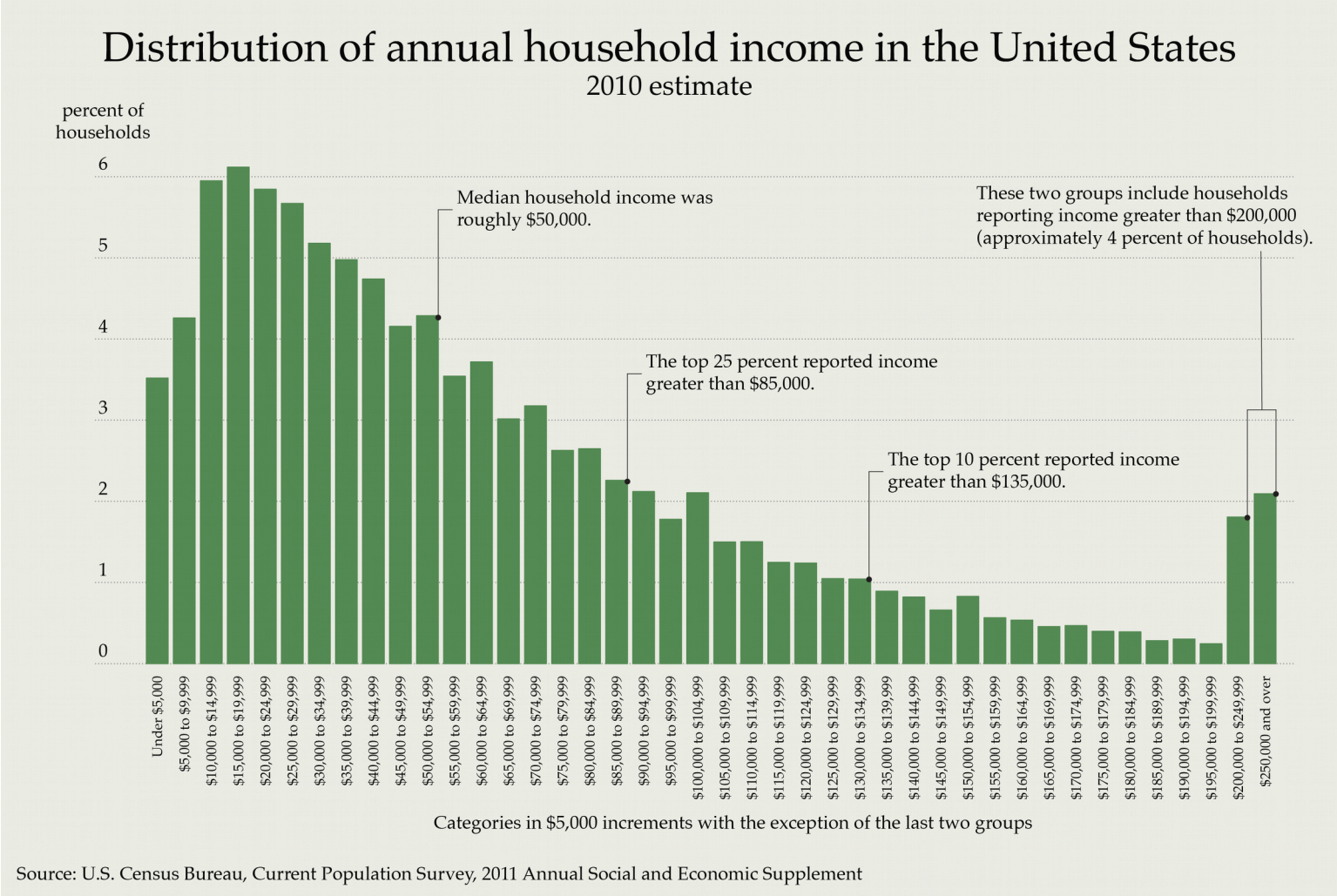
Region type is similar in that most "suburban" category is a dominant one, but it's too broad of a category to give us much insight. Additionally, the "t" and "s" categories are assumed to reference the "town" and "suburban" categories and thus shouldn't be separate.

Marriage status has the majority in "unknown" category, this makes the usefulness of this statistic questionable.

Income plot has many values in the zero category. At a glance it appears unclear whether "0" is high or low-income category. However, if the categories were to be rearranged from 9 to 0, the picture would look like a bell curve, that is skewed right and has high values for "0" category to the right. This may seem odd at first, but my research shows that this distribution very much resembles that of a typical US income distribution. **As in the picture below titled *"Distribution of annual household income in the United States".***

Also, after I inspected the category "home owner" as a rough measure of wealth, I saw that most of the customers in category "9" were not home owners. As I continued through categories from 9 to 8 to 7 etc. I found that the home ownership increased. When I reached category "0", most of the customers were home owners.

Therefore, my initial conclusion is that the "0" category isn't an outlier and should be assumed typical. Furthermore, the income category distribution is ascending form "9" to "0", where 9 represents low-income and 0 represents high-income. To again, clarify, while the category values themselves are descending, the income amount for each category is ascending.



[source: https://jerclifton.files.wordpress.com/2012/10/distribution_of_annual_household_income_in_the_united_states.png]

Credit rating is dominated by "B" category.

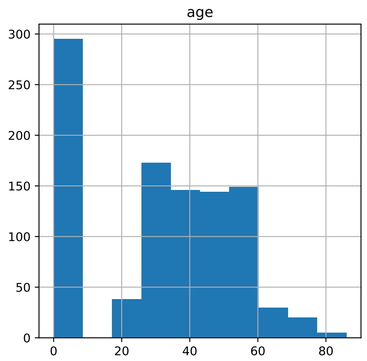
Credit card shows us most people have credit cards. Also, there are extra categories "t", "f", "yes" and "no", which carry the same meaning as "true" and "false" categories and are thus unnecessary.

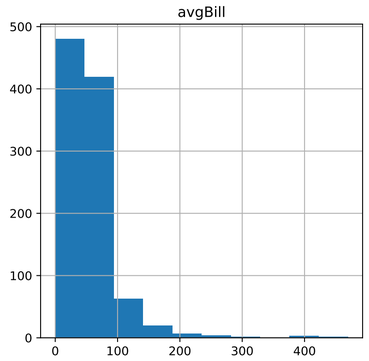
Children plot indicates that majority of customers do not have children.

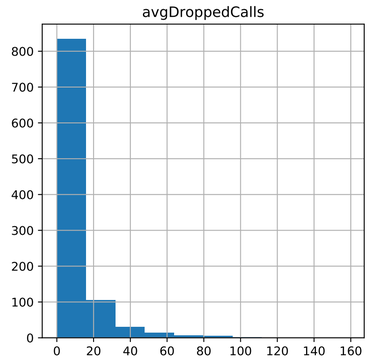
Smartphone plot shows us that majority of customers have smartphones.

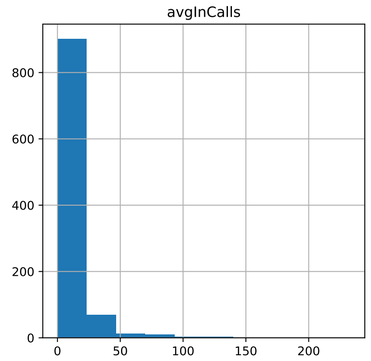
Home owner plot indicates that about two thirds of customers ore not home owners.

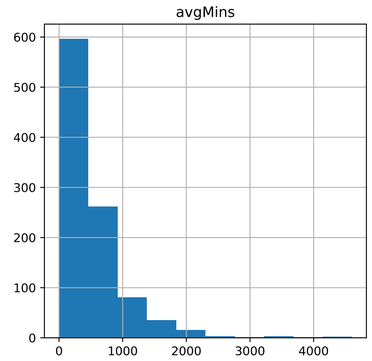
Churn plot shows us that customers are roughly equally likely to churn as they are to not churn.

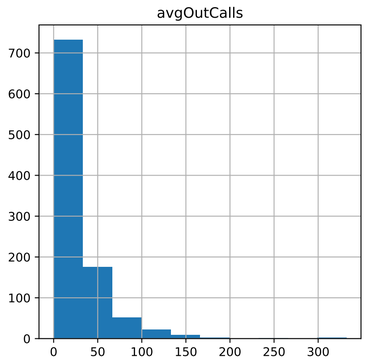

age

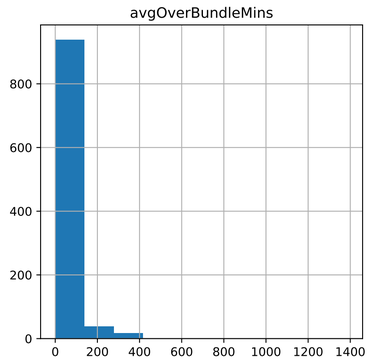

avgBill

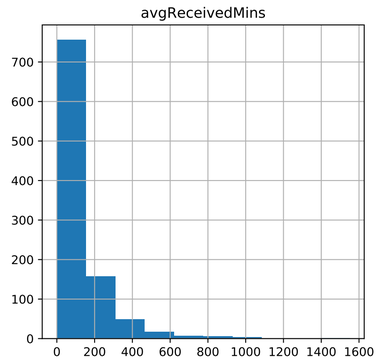

avgDroppedCalls

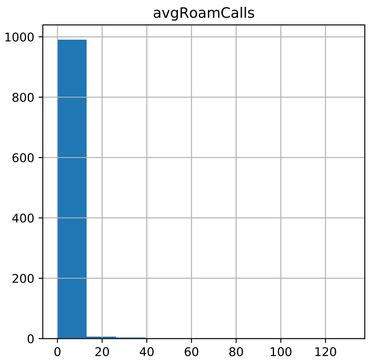

avgInCalls

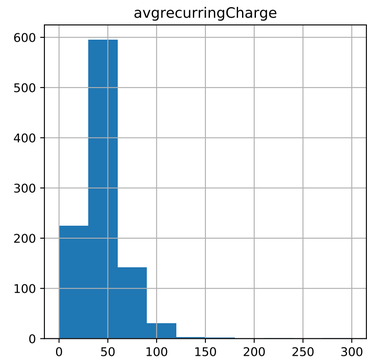

avgMins

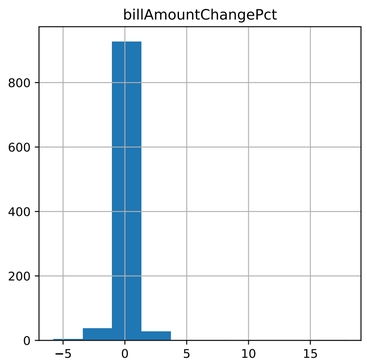

avgOutCalls

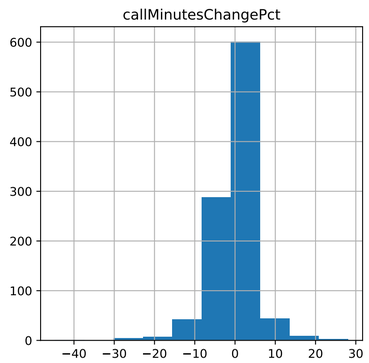

avgOverBundleMins

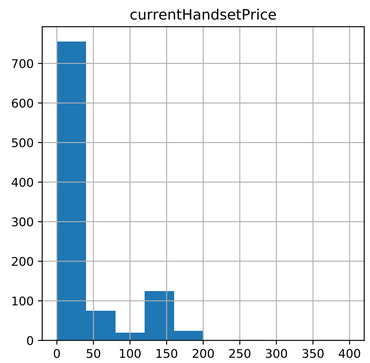

avgReceivedMins

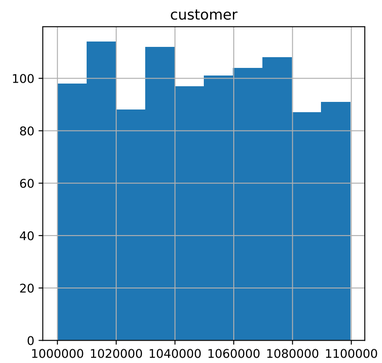

avgRoamCalls

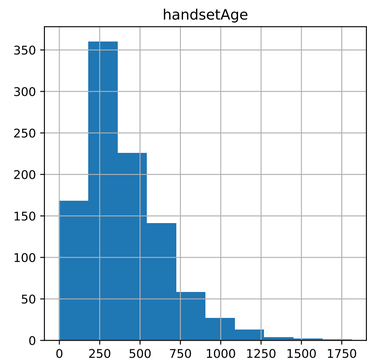

avgrecurringCharge

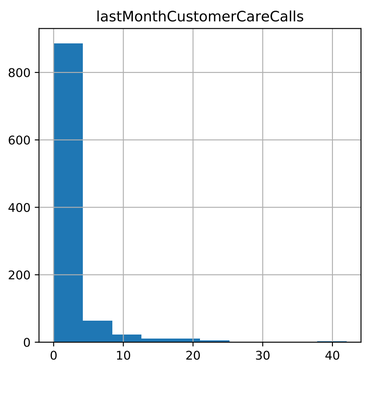

billAmountChangePct

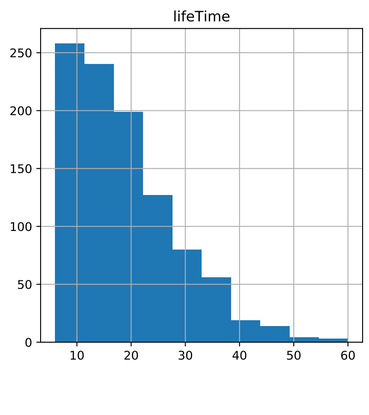

callMinutesChangePct

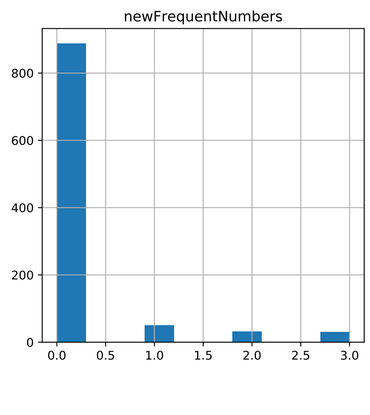

currentHandsetPrice

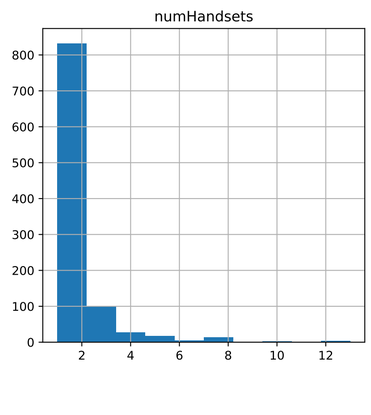

customer

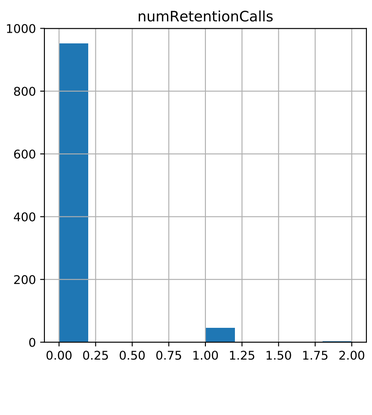

handsetAge

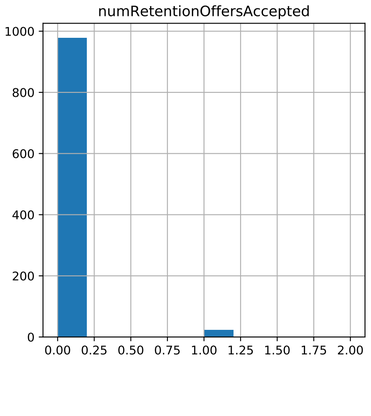

lastMonthCustomerCareCalls

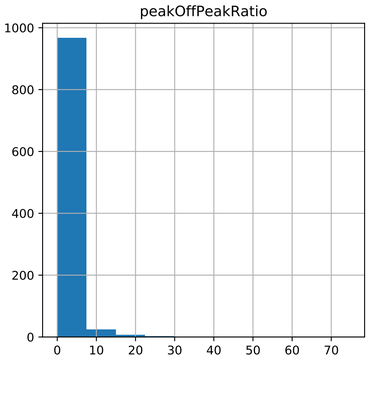

lifeTime

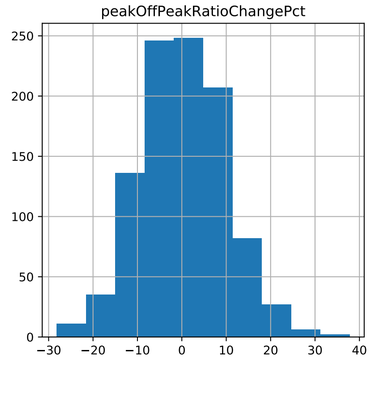

newFrequentNumbers

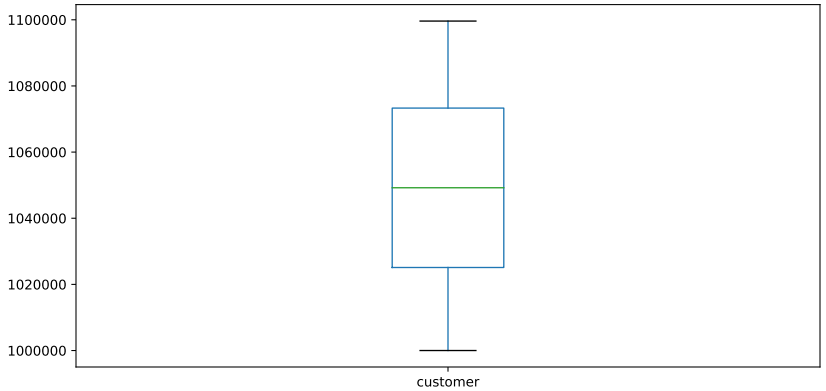

numHandsets

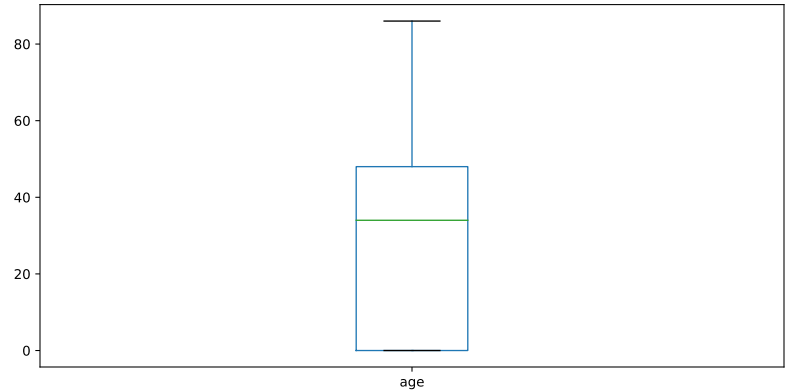

numRetentionCalls

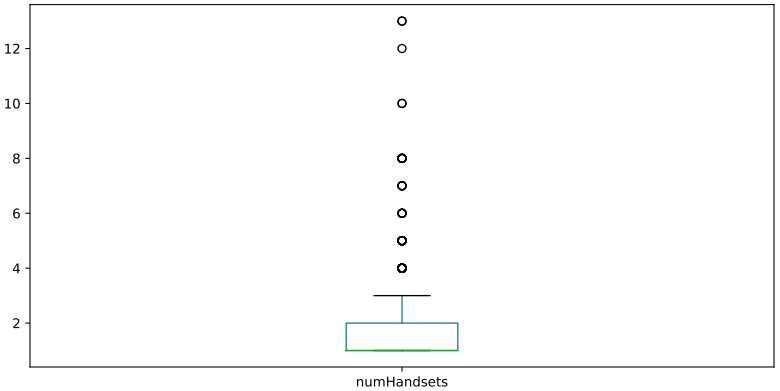

numRetentionOffersAccepted

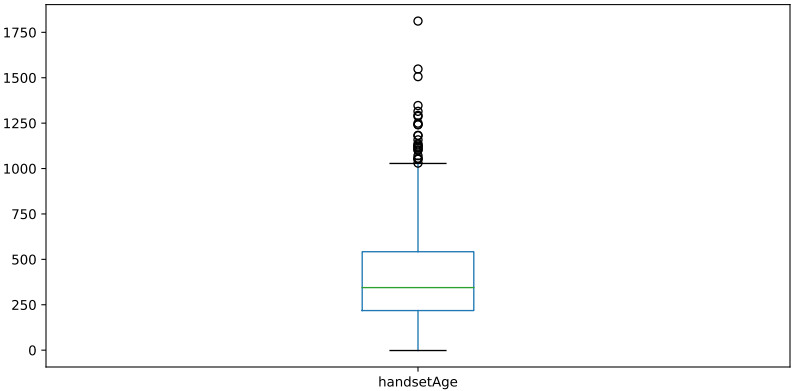

peakOffPeakRatio

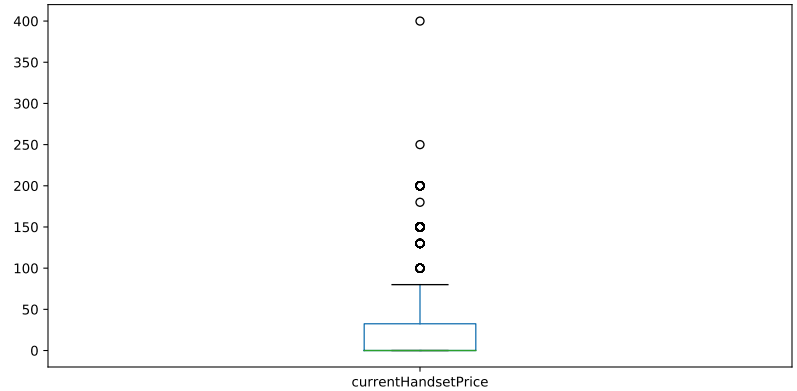

peakOffPeakRatioChangePct

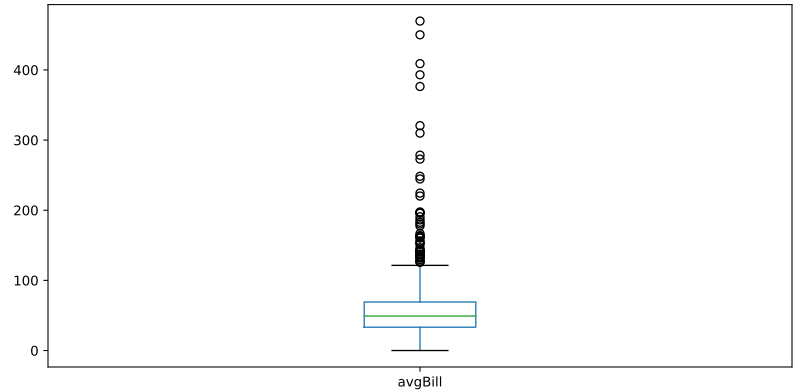

1100000
1080000
1060000
1040000
1020000
1000000
customer

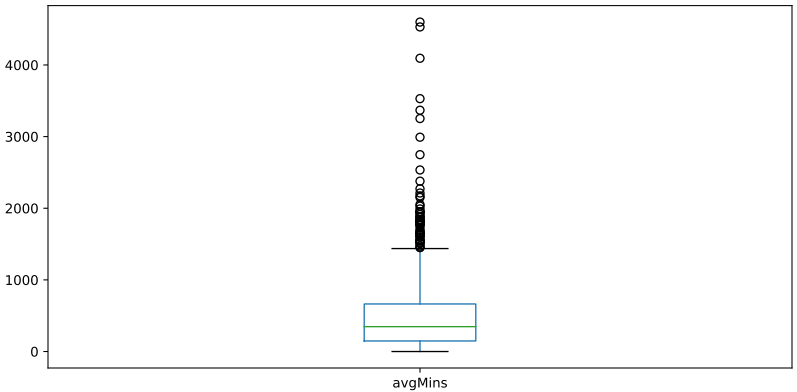

80
60
40
20
0
age

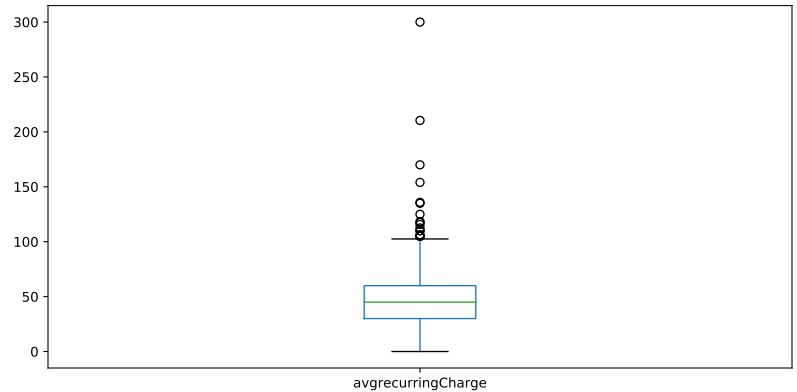

12
10
8
6
4
2
numHandsets

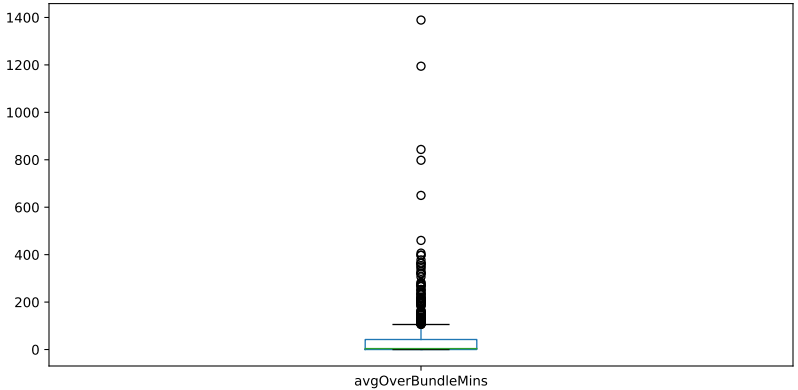

1750
1500
1250
1000
750
500
250
0
handsetAge

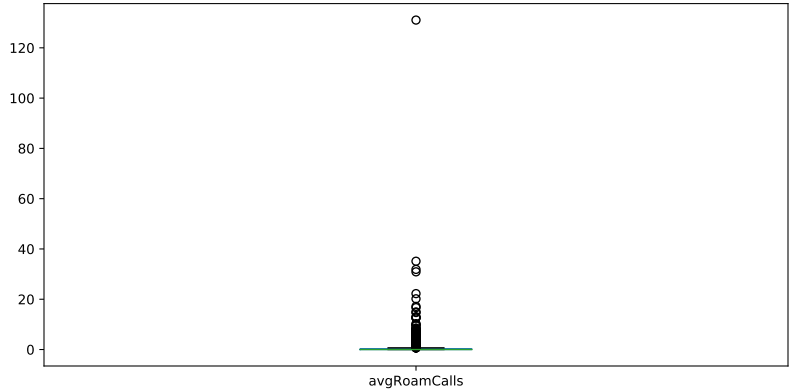

400
350
300
250
200
150
100
50
0
currentHandsetPrice

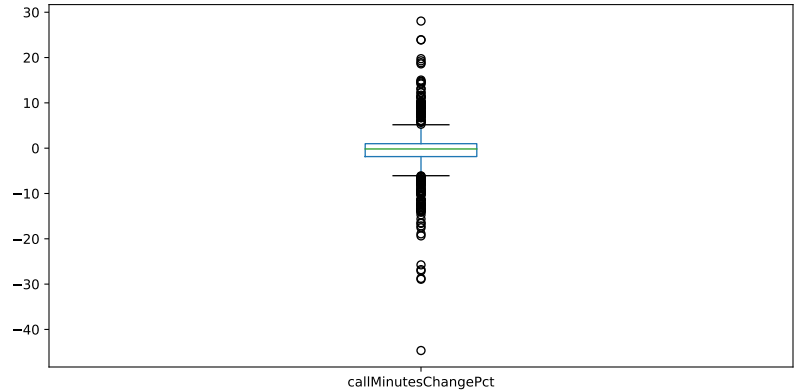

400
300
200
100
0
avgBill

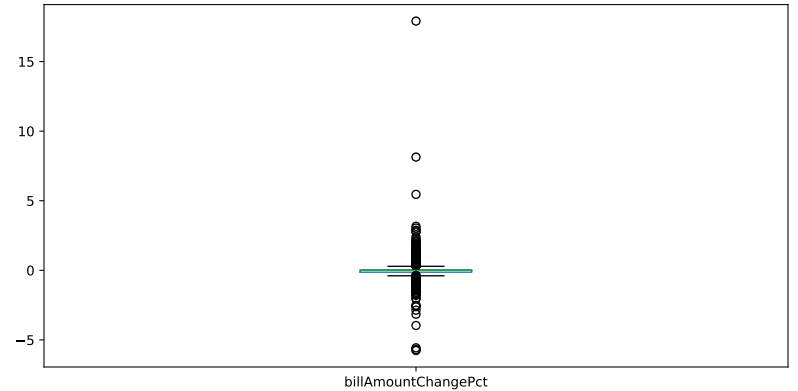

4000
3000
2000
1000
0
avgMins

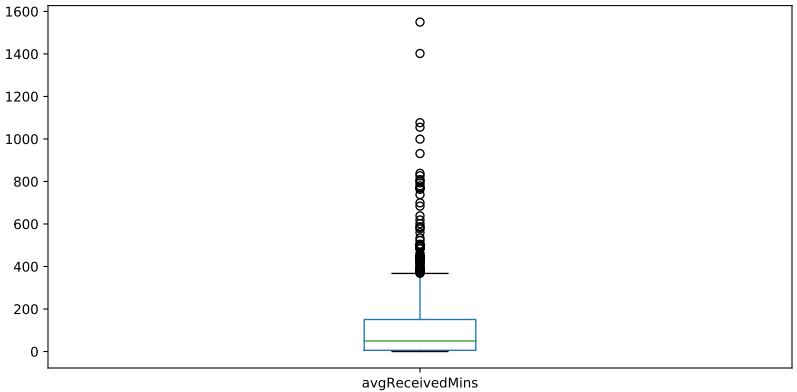

300
250
200
150
100
50
0
avgrecurringCharge

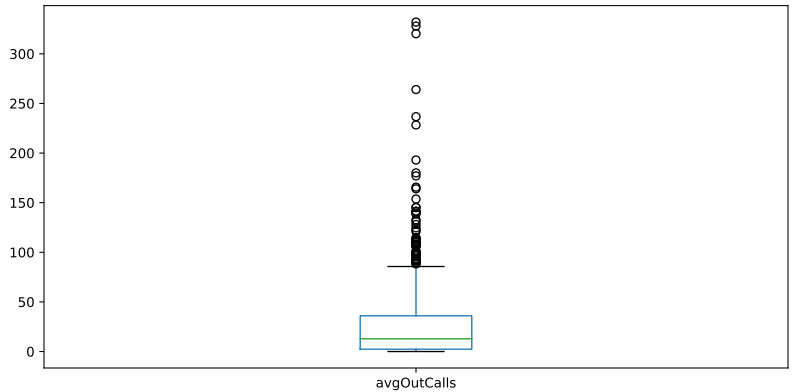

1400
1200
1000
800
600
400
200
0
avgOverBundleMins

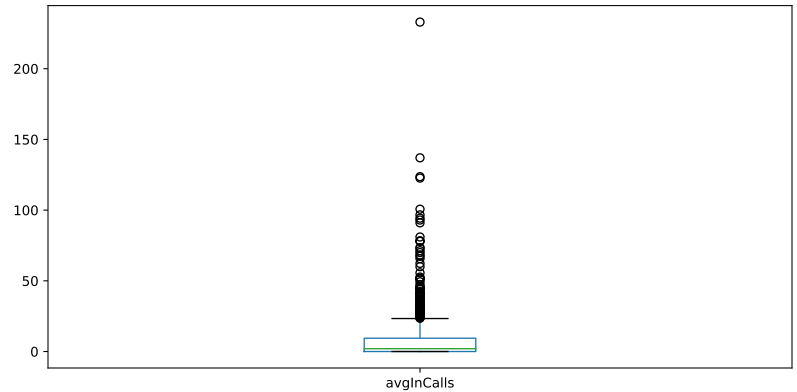

120
100
80
60
40
20
0
avgRoamCalls

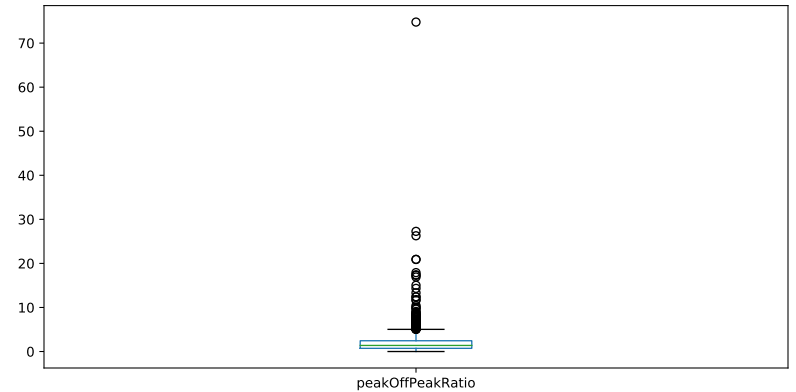

30
20
10
0
−10
−20
−30
−40
callMinutesChangePct

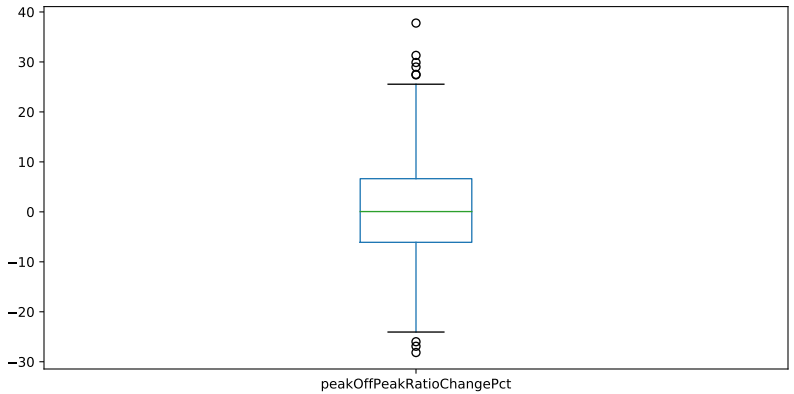

15
10
5
0
−5
billAmountChangePct

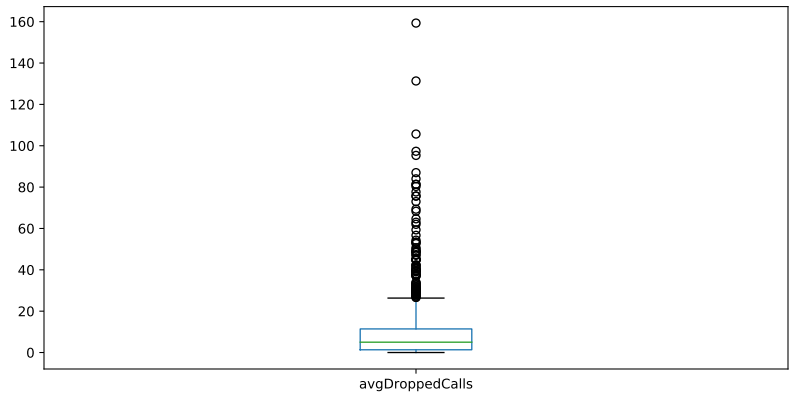

1600
1400
1200
1000
800
600
400
200
0
avgReceivedMins

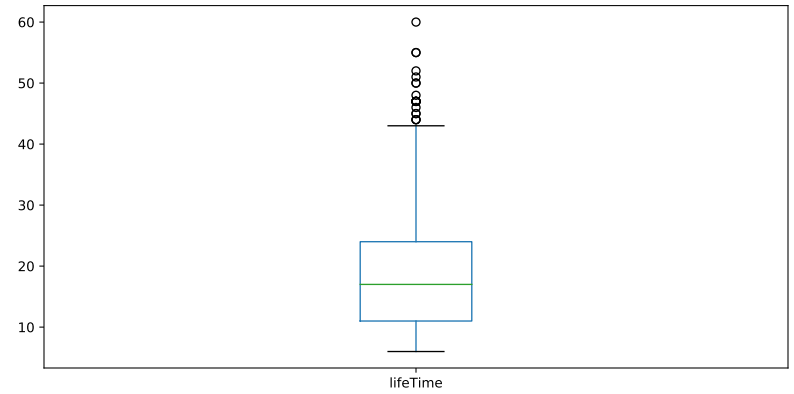

300
250
200
150
100
50
0
avgOutCalls


200
150
100
50
0
avgInCalls


70
60
50
40
30
20
10
0
peakOffPeakRatio

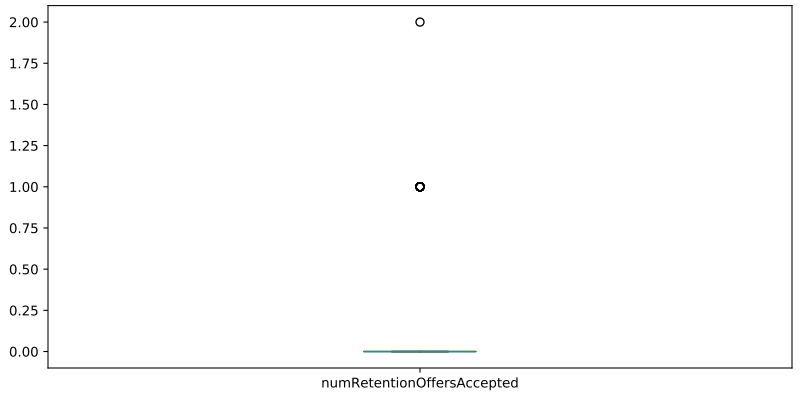

40
30
20
10
0
−10
−20
−30
peakOffPeakRatioChangePct

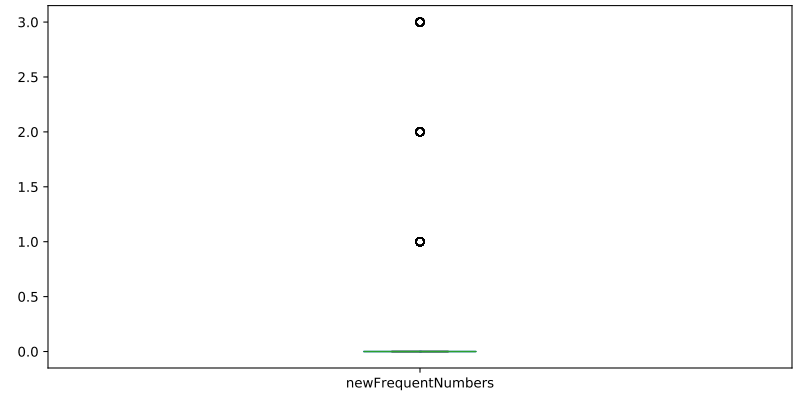

160
140
120
100
80
60
40
20
0
avgDroppedCalls

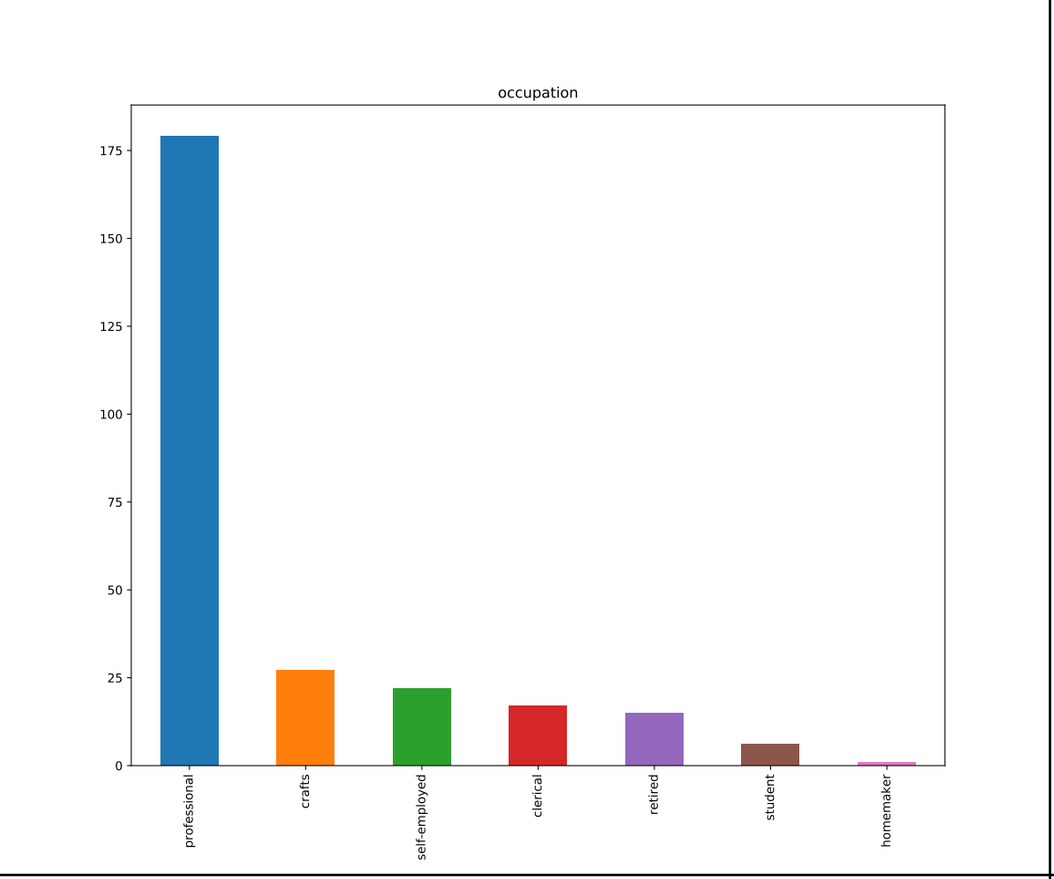

60
50
40
30
20
10
lifeTime

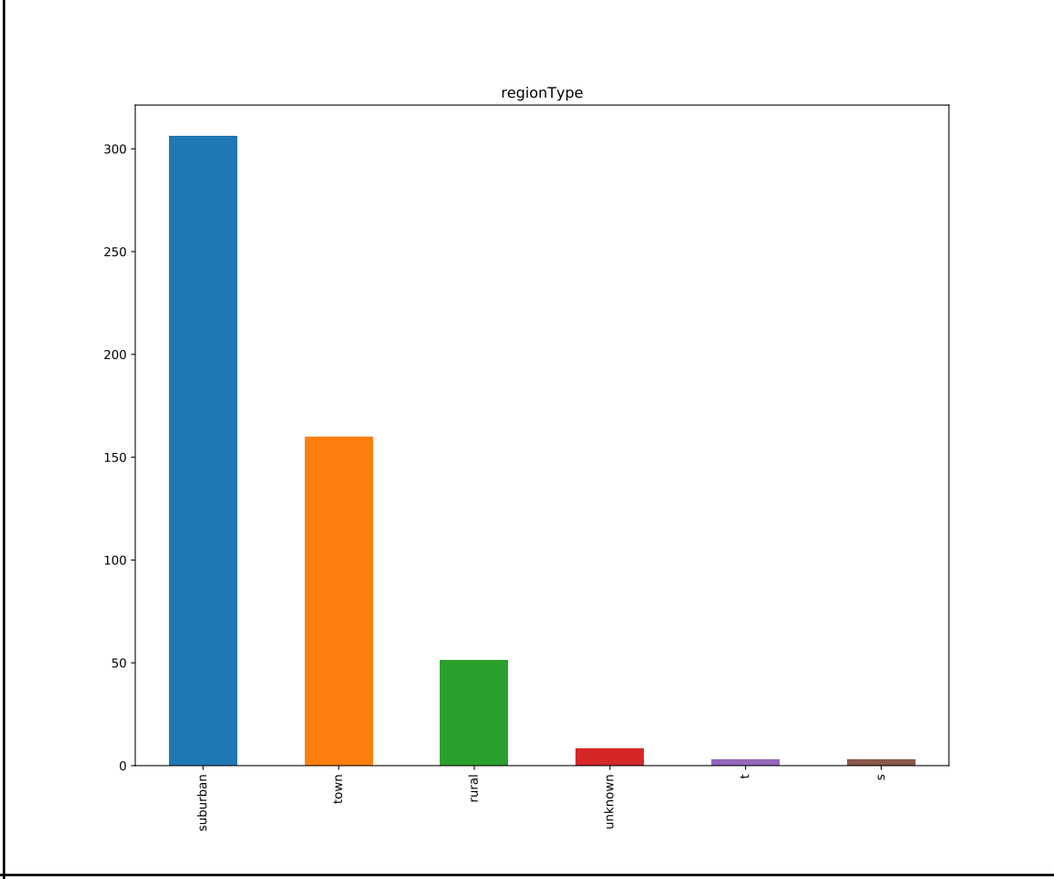

40
30
20
10
0
lastMonthCustomerCareCalls

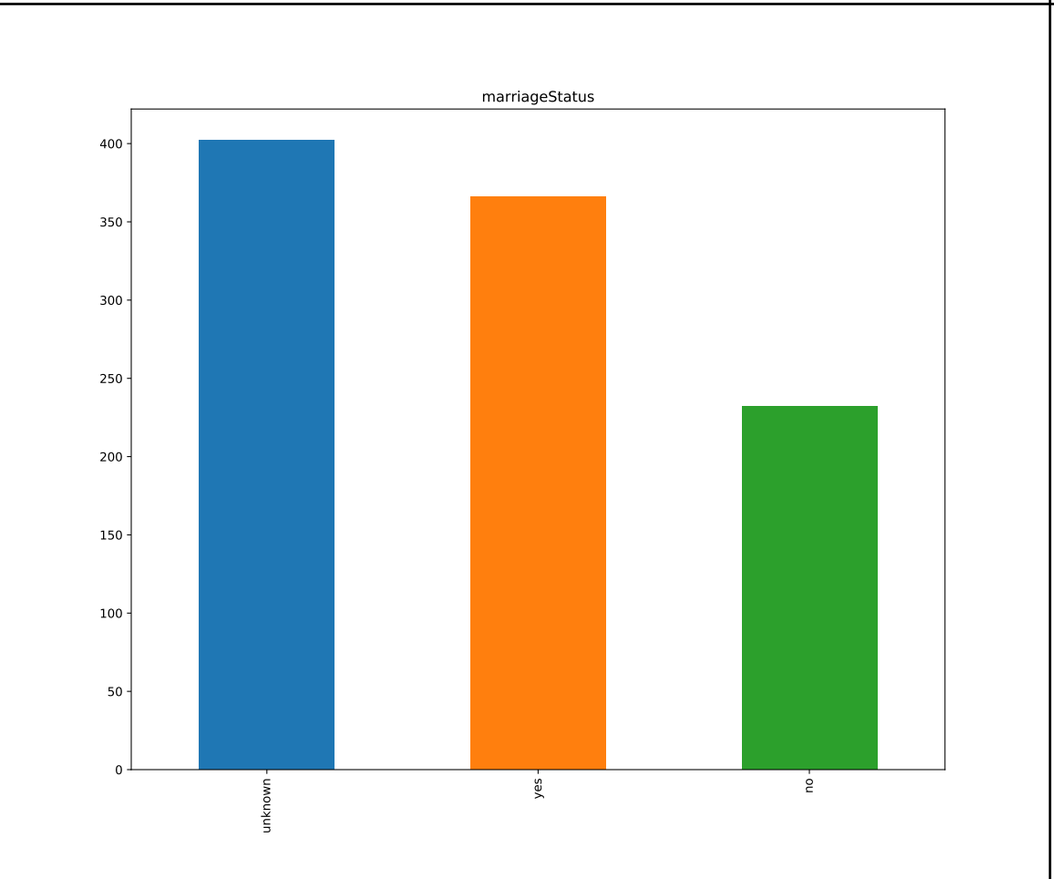

2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
numRetentionCalls

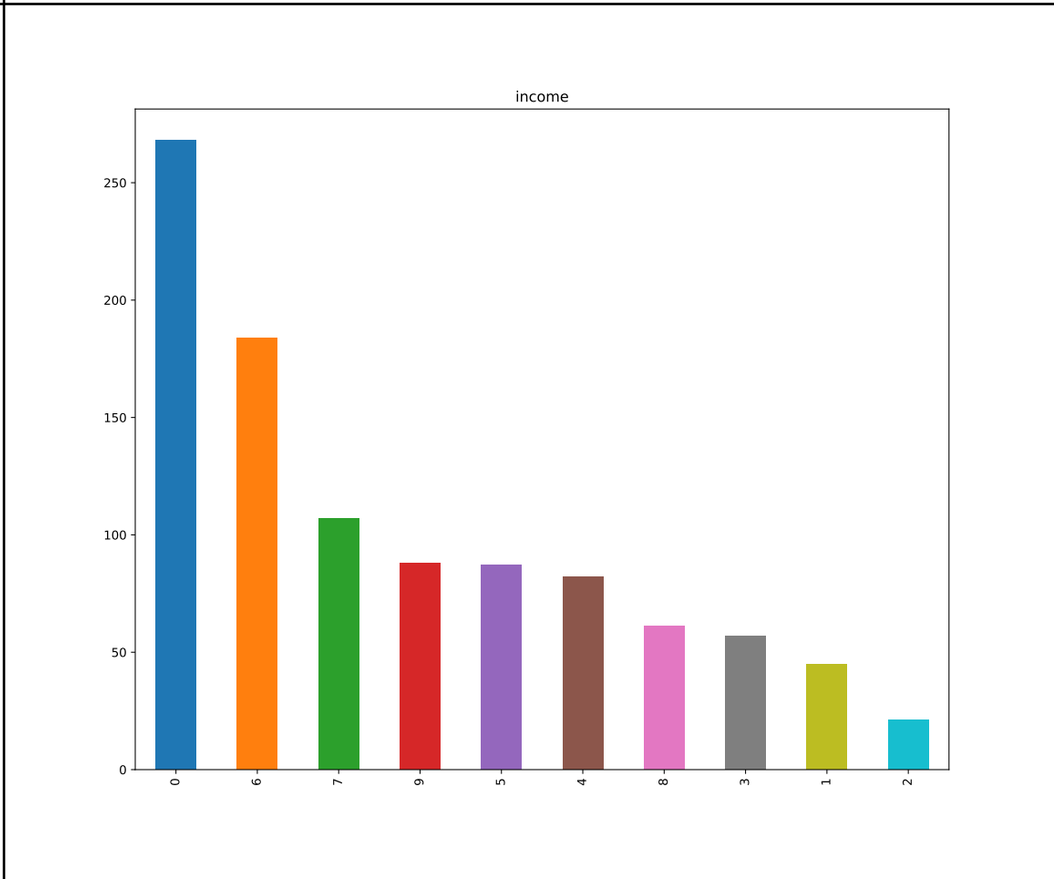

2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
numRetentionOffersAccepted

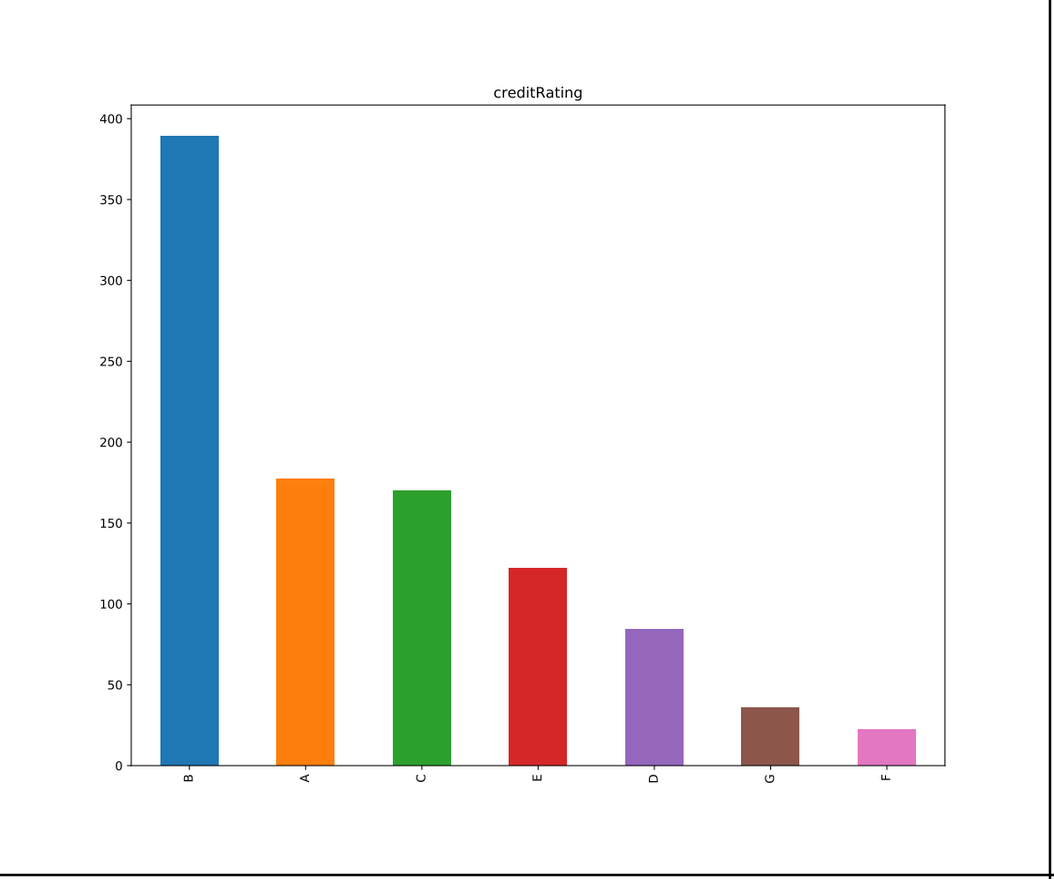

3.0
2.5
2.0
1.5
1.0
0.5
0.0
newFrequentNumbers

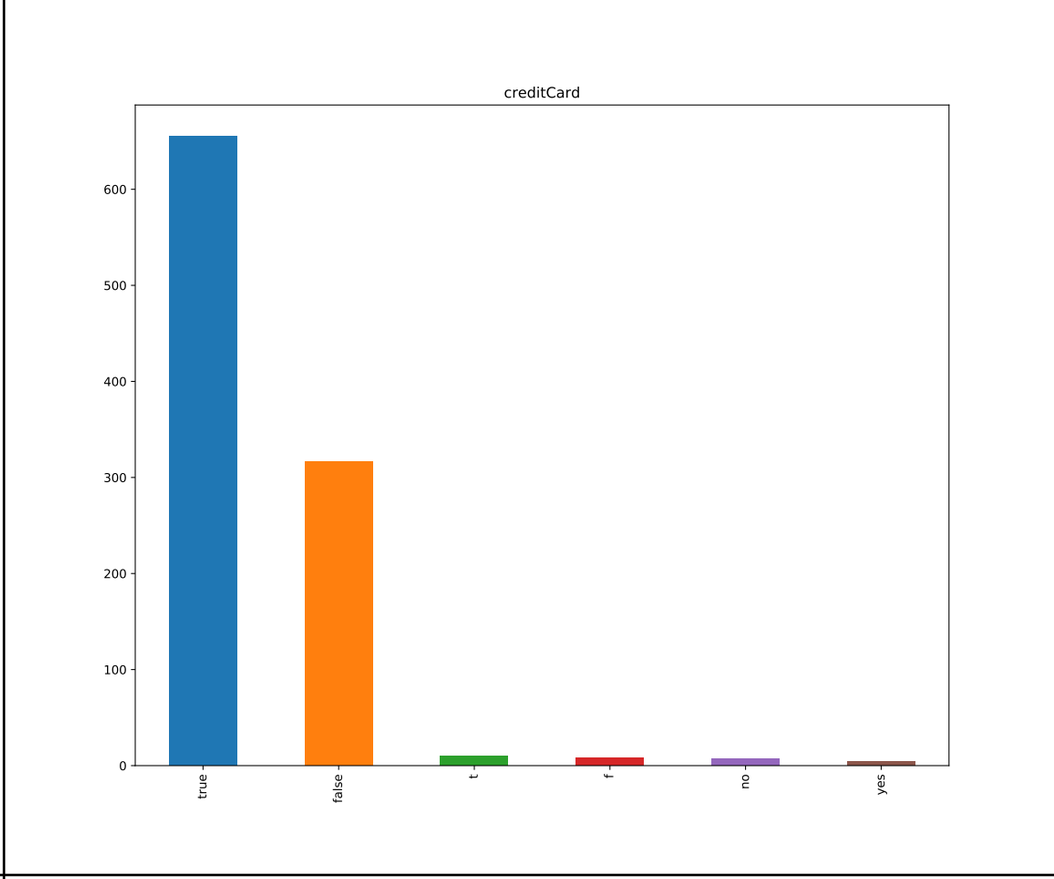

occupation
175
150
125
100
75
50
25
0
professional
crafts
self-employed
clerical
retired
student
homemaker

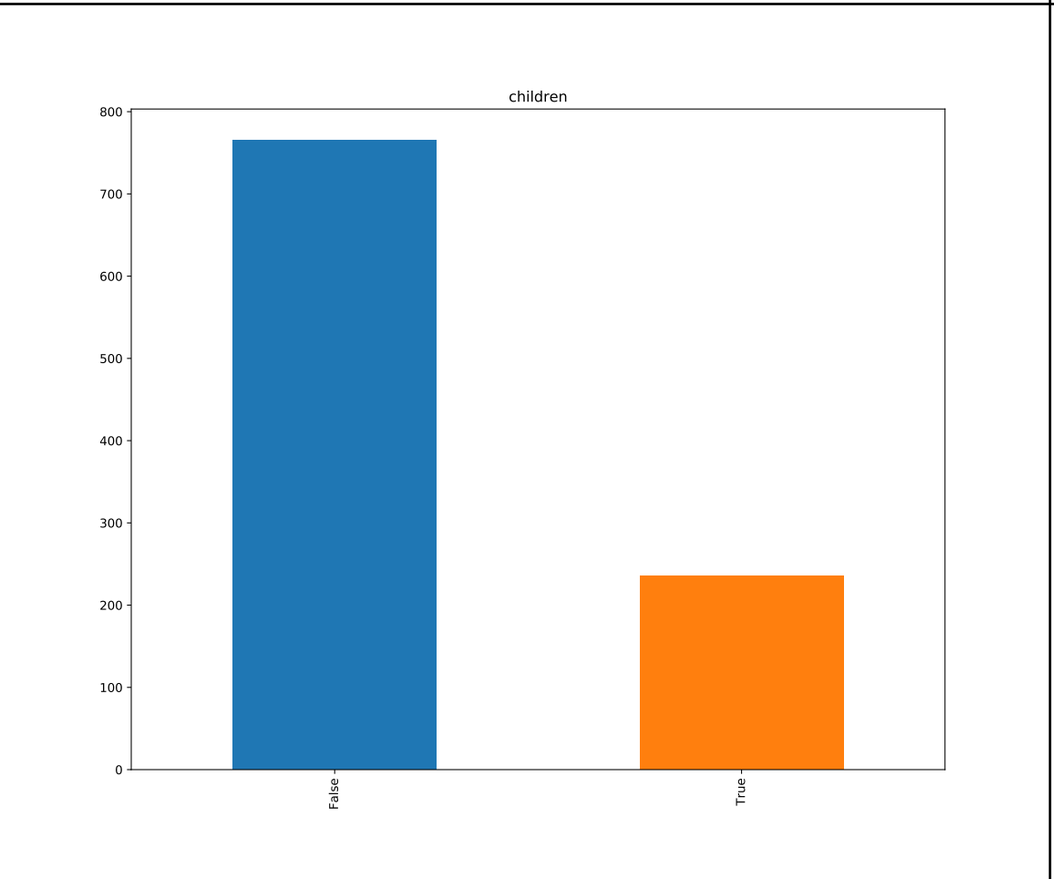

regionType
300
250
200
150
100
50
0
suburban
town
rural
unknown
t
s

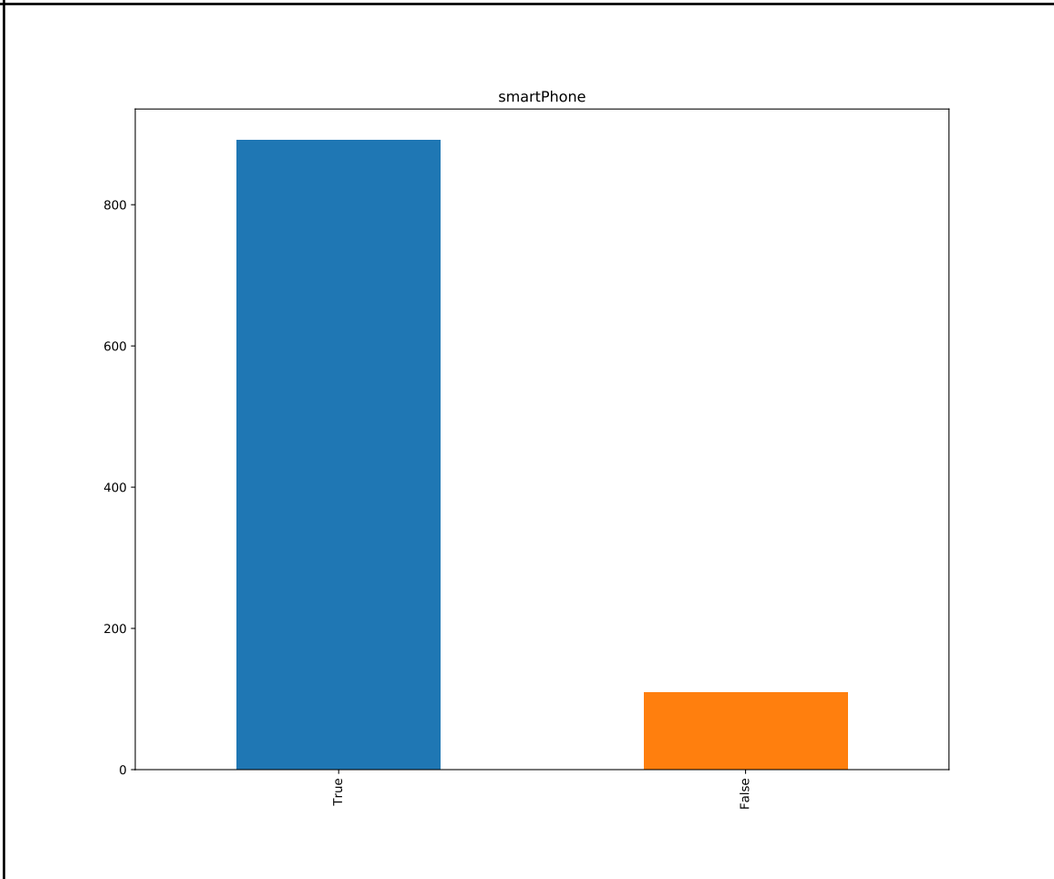

marriageStatus
400
350
300
250
200
150
100
50
0
unknown
yes
no

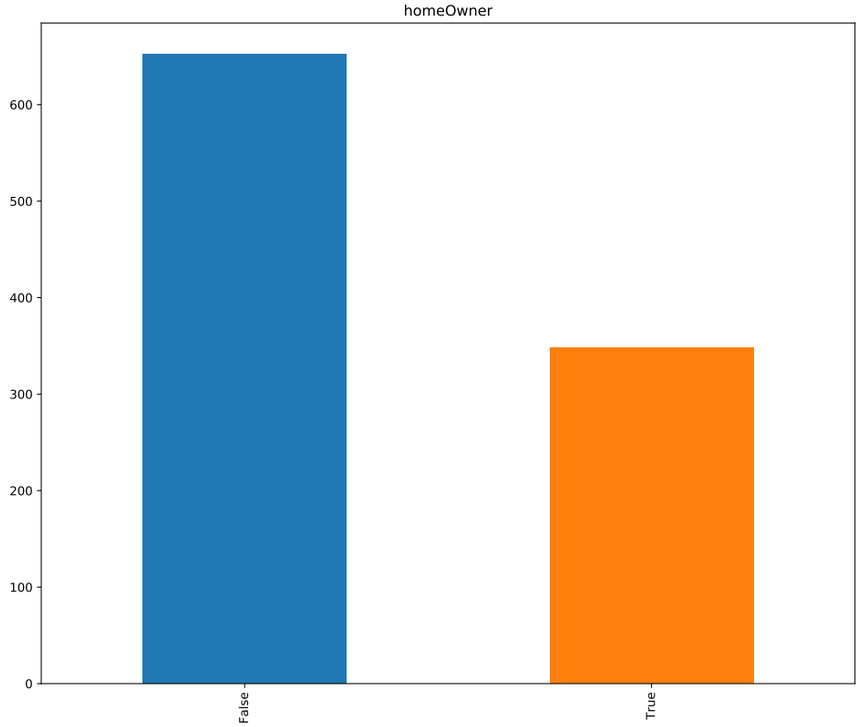

income
250
200
150
100
50
0
0
6
7
9
5
4
8
3
1
2

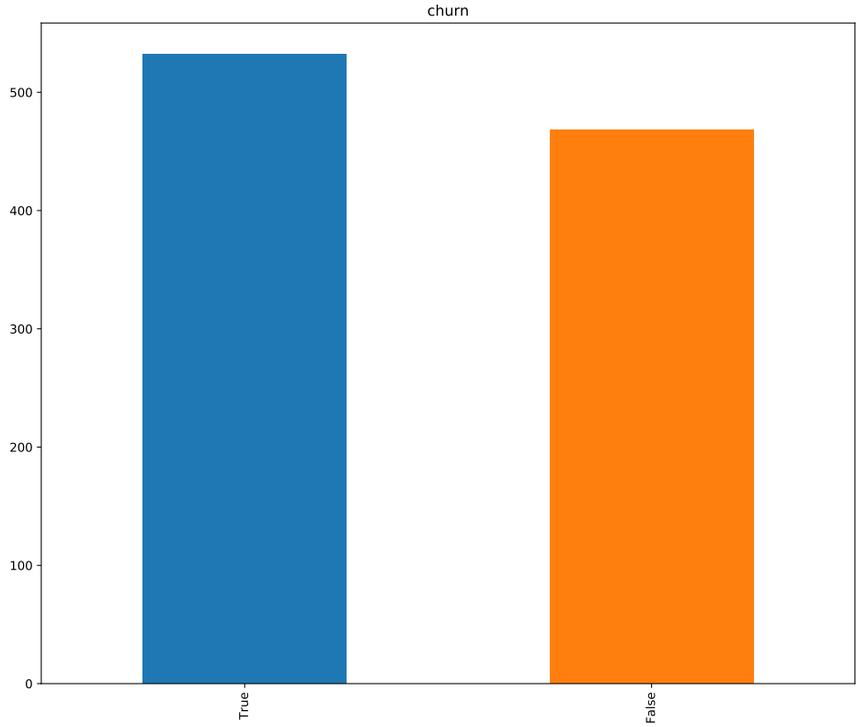

creditRating
400
350
300
250
200
150
100
50
0
B
A
C
E
D
G
F


creditCard
600
500
400
300
200
100
0
true
false
t
f
no
yes


children
800
700
600
500
400
300
200
100
0
False
True


smartPhone
800
600
400
200
0
True
False


homeOwner
600
500
400
300
200
100
0
False
True


churn
500
400
300
200
100
0
True
False