Axis Insurance – Project 2

Sarah A. Thomas

Background

Axis Insurance has tasked their Analytics department with a small statistical study to determine what factors may impact insurance claims.

Objectives

- Using Exploratory Data Analysis, find insights about the given dataset detailing policy holder information.
- Perform statistical tests for the following claims (create null hypotheses and corresponding alternative hypotheses):
  - Prove (or disprove) that the medical claims made by the people who smoke is greater than those who don’t.
  - Prove (or disprove) with statistical evidence that the BMI of females is different from that of males.
  - Find whether the proportion of smokers is significantly different across different regions.
  - Find whether the mean BMI of women with no children, one child, and two children are the same.
  - *Consider a significance level of 0.05 for all tests.

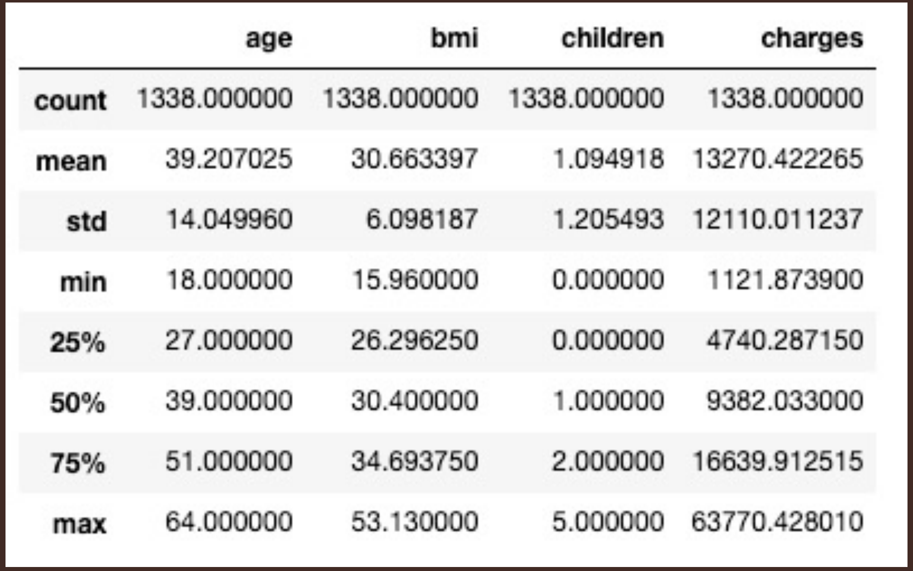
Data Information (copied from Data Dictionary)

Variable	Description
Age	This is an integer indicating the age of the primary beneficiary (excluding those above 64 years, since they are generally covered by the government).
Sex	This is the policy holder's gender, either male or female.
BMI	This is the body mass index (BMI), which provides a sense of how over or underweight a person is relative to their height. BMI is equal to weight (in kilograms) divided by height (in meters) squared. An ideal BMI is within the range of 18.5 to 24.9.
Children	This is an integer indicating the number of children/dependents covered by the insurance plan.
Smoker	This is yes or no depending on whether the insured regularly smokes tobacco.
Region	This is the beneficiary's place of residence in the U.S., divided into four geographic regions - northeast, southeast, southwest, or northwest.
Charges	Individual medical costs billed to health insurance.

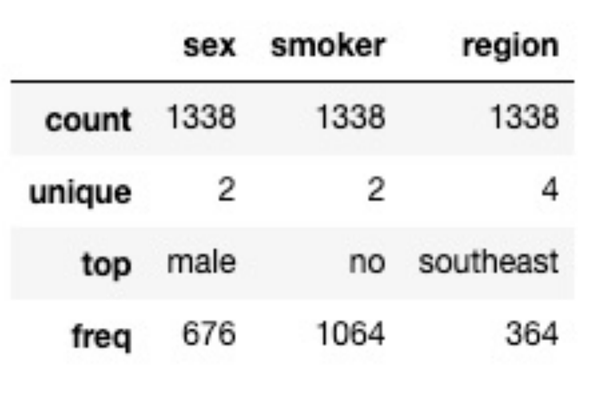
Shape of the Data: 1338 rows, 7 columns

Exploratory Data Analysis – Initial Observations

- Age ranges from 18-64, with mean and median very close in value (mean = 39.207, median = 39.000). This indicates near zero skewness.
- BMI ranges from 15.96-53.13, with mean and median very close in value (mean = 30.663, median = 30.400). This indicates near zero skewness.
- Number of children ranges from 0-5 with mean and median very close in value (mean = 1.095, median = 1.000). This indicates near zero skewness.
- Charges range from 1121.87-63770.43, a wide range. With the mean (13,270.42) greater than the median (9,382.03), the data is right-skewed.
- More males are policy holders (676) compared to females (662).
- Most policy holders do not smoke (1064).
- Most policy holders live in the southeast region of the U.S. (364).



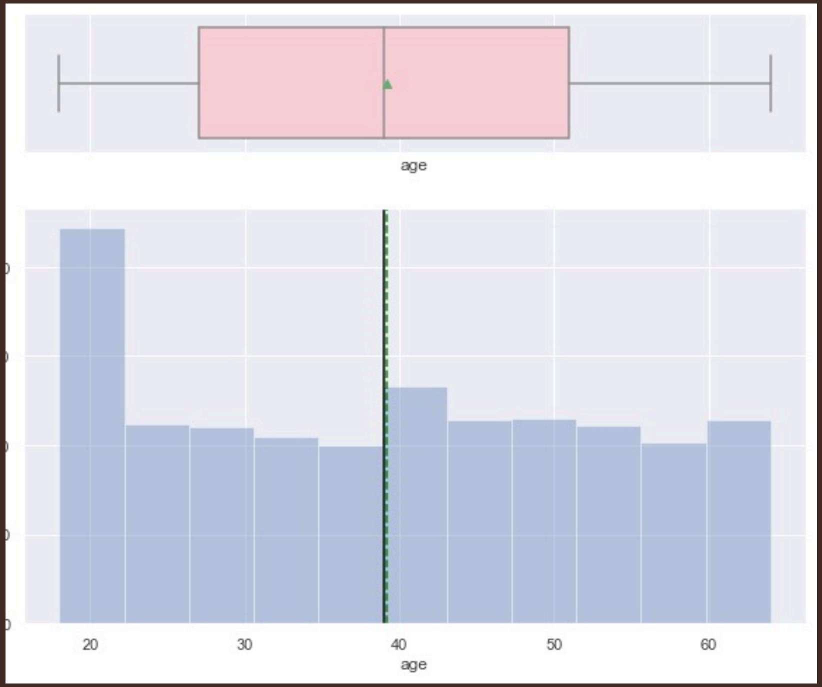
	age	bmi	children	charges
count	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.207025	30.663397	1.094918	13270.422265
std	14.049960	6.098187	1.205493	12110.011237
min	18.000000	15.960000	0.000000	1121.873900
25%	27.000000	26.296250	0.000000	4740.287150
50%	39.000000	30.400000	1.000000	9382.033000
75%	51.000000	34.693750	2.000000	16639.912515
max	64.000000	53.130000	5.000000	63770.428010



	sex	smoker	region
count	1338	1338	1338
unique	2	2	4
top	male	no	southeast
freq	676	1064	364

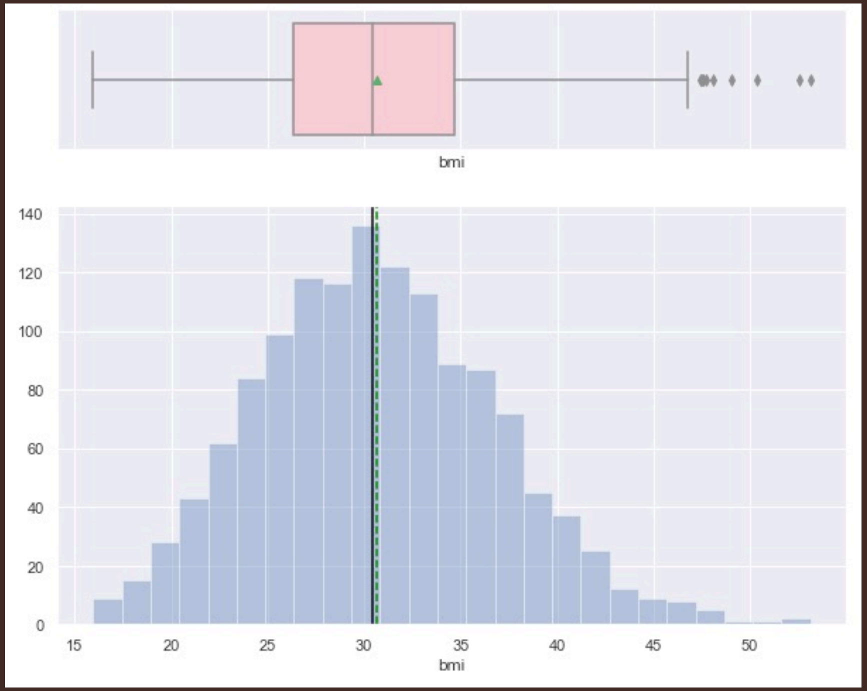
Univariate Analysis (1 of 3)

Age



- There are no outliers.
- Mean and median are (approx.) 39 years.
- Q3 is 51 which means that 75% of customers are below age 51.

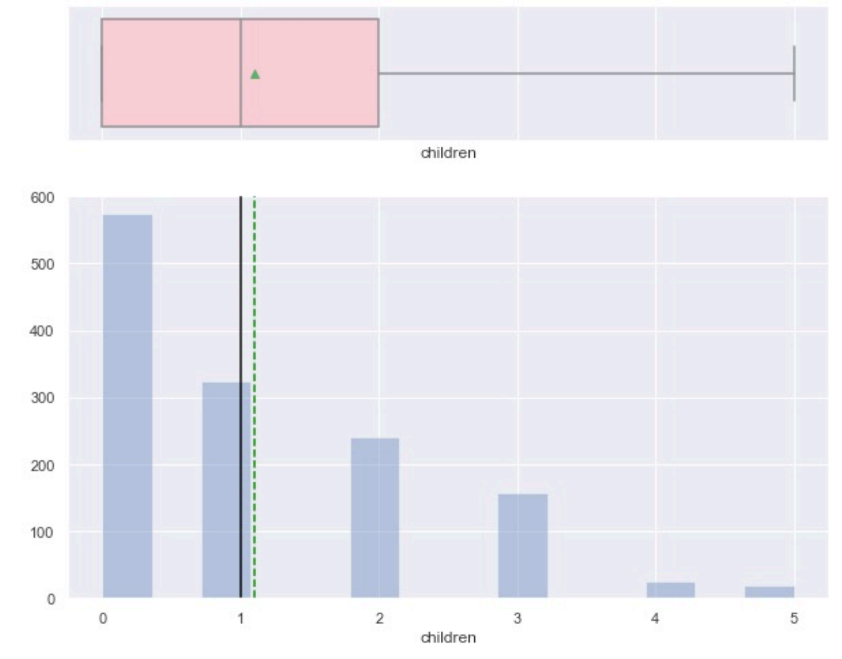
BMI



- There are outliers for this variable.
- The mean and median are (approx.) 31.
- Q3 is 35 which means that 75% of customers have a BMI below 35.

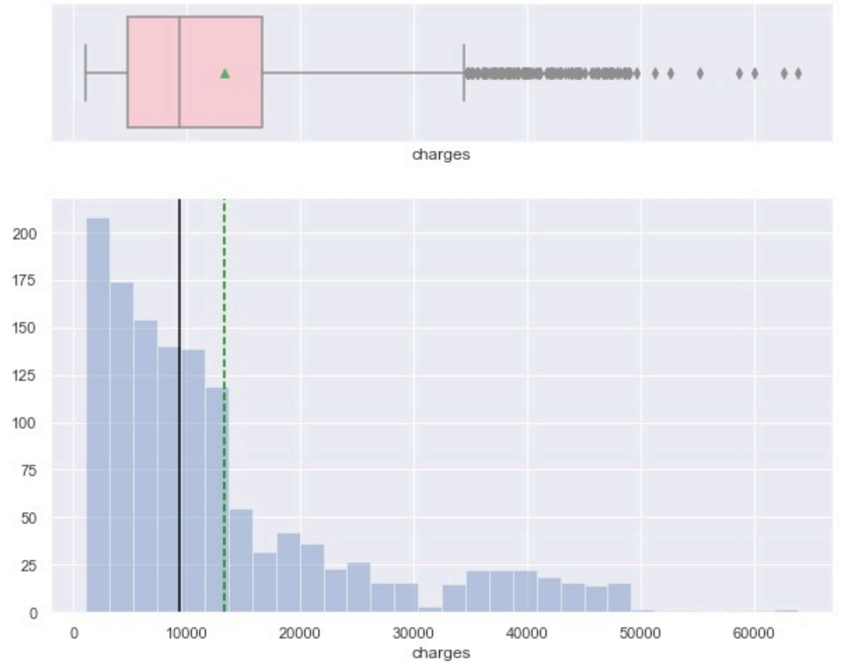
Univariate Analysis (2 of 3)

Children



- There are no outliers.
- The mean and median are (approx.) 1 child.
- Q3 is 2 which means that 75% of policy holders have less than 2 children.

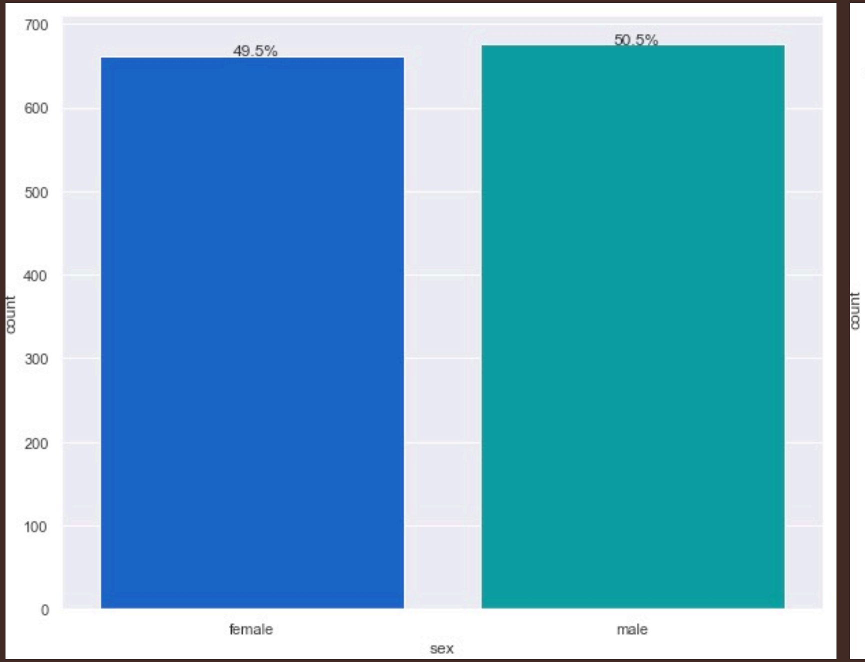
Charges



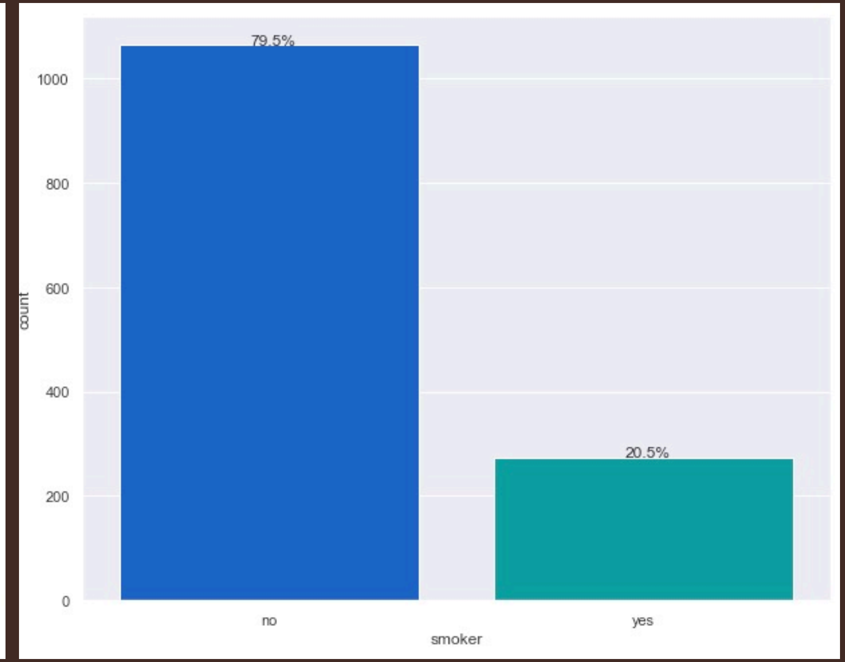
- There are outliers for this variable all on the higher end (above 35,000).
- The mean is greater than the median, therefore the data is right-skewed.
- Q3 is approx. 17,000. 75% of customer have been charged less than the value of Q3.

Univariate Analysis (3 of 3)

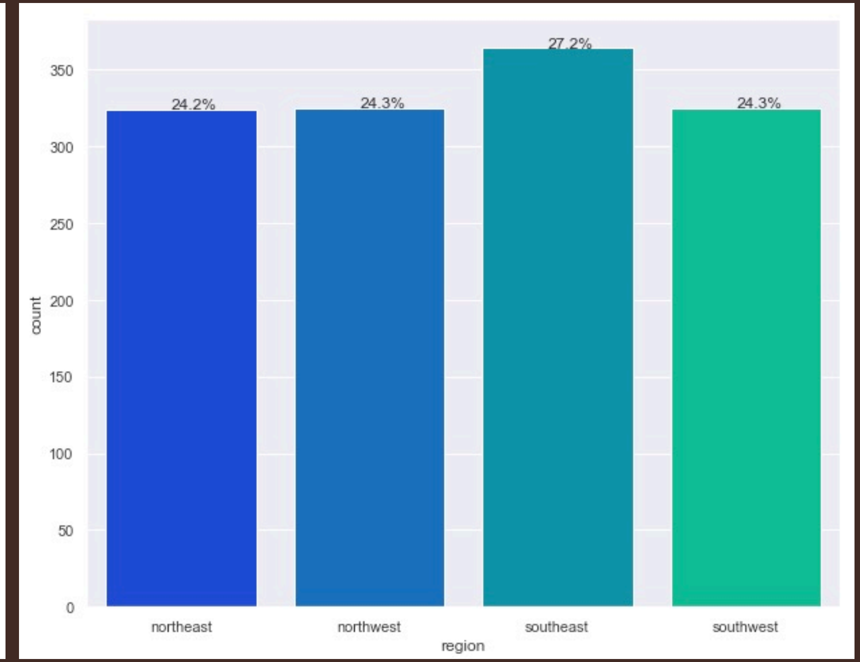
Sex



Smoker



Region



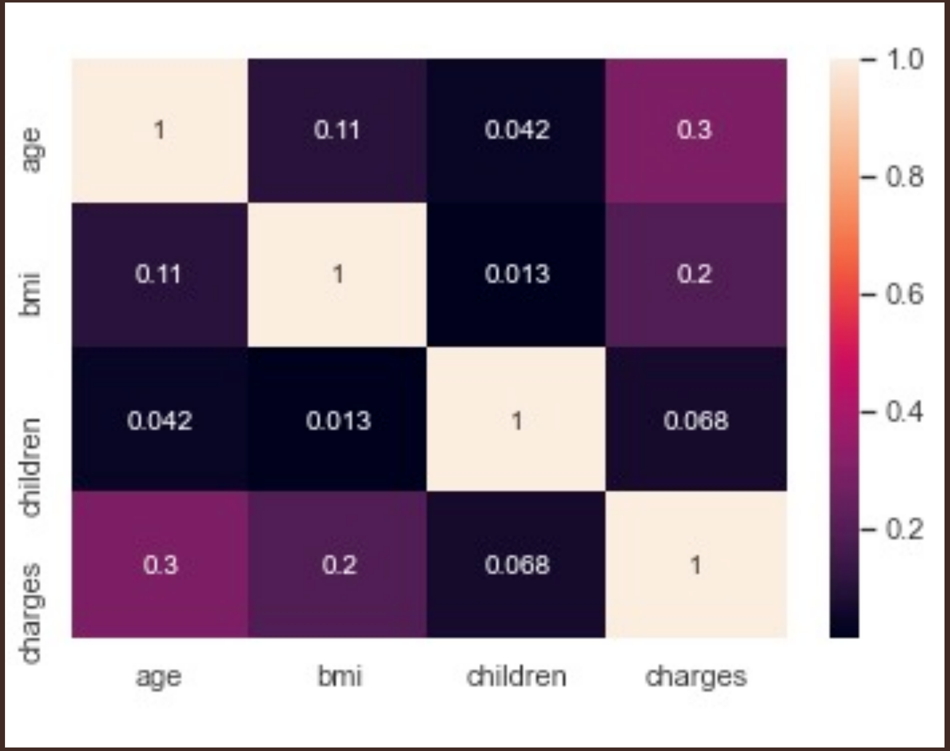
There are more male than female policy holders but just barely by 1% (49.5% vs 50.5%).

Non-smokers greatly outnumber smokers (79.5% vs. 20.5%).

More policy holders live in the southeast region than the other regions, but the rest of the policy holders are evenly split between the remaining three regions.

Correlation

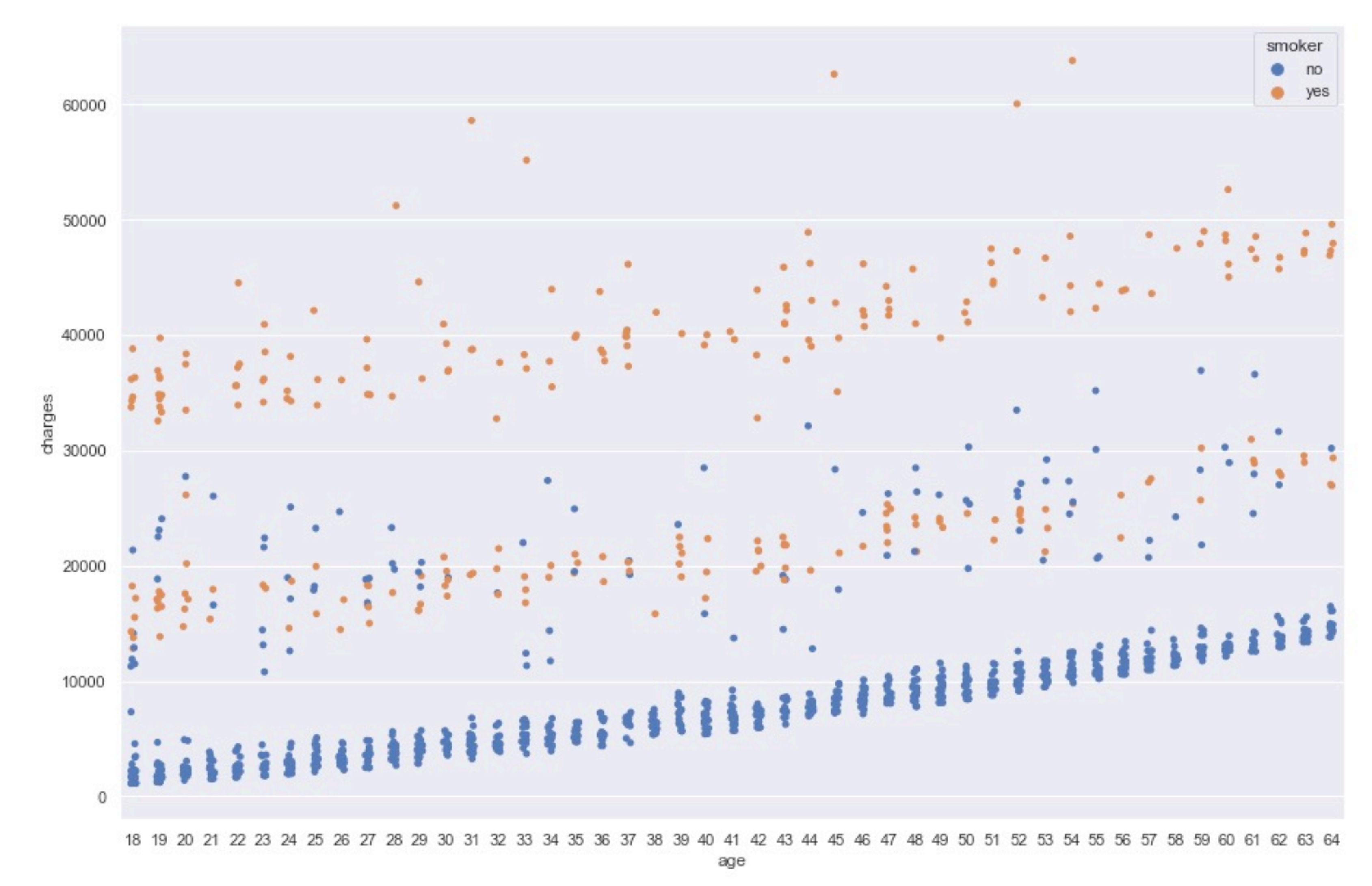
Correlations worth investigating in a multivariate analysis include age vs. charges and bmi vs. charges.



Multivariate Analysis (1 of 2)

Age vs. Charges vs. Smoker

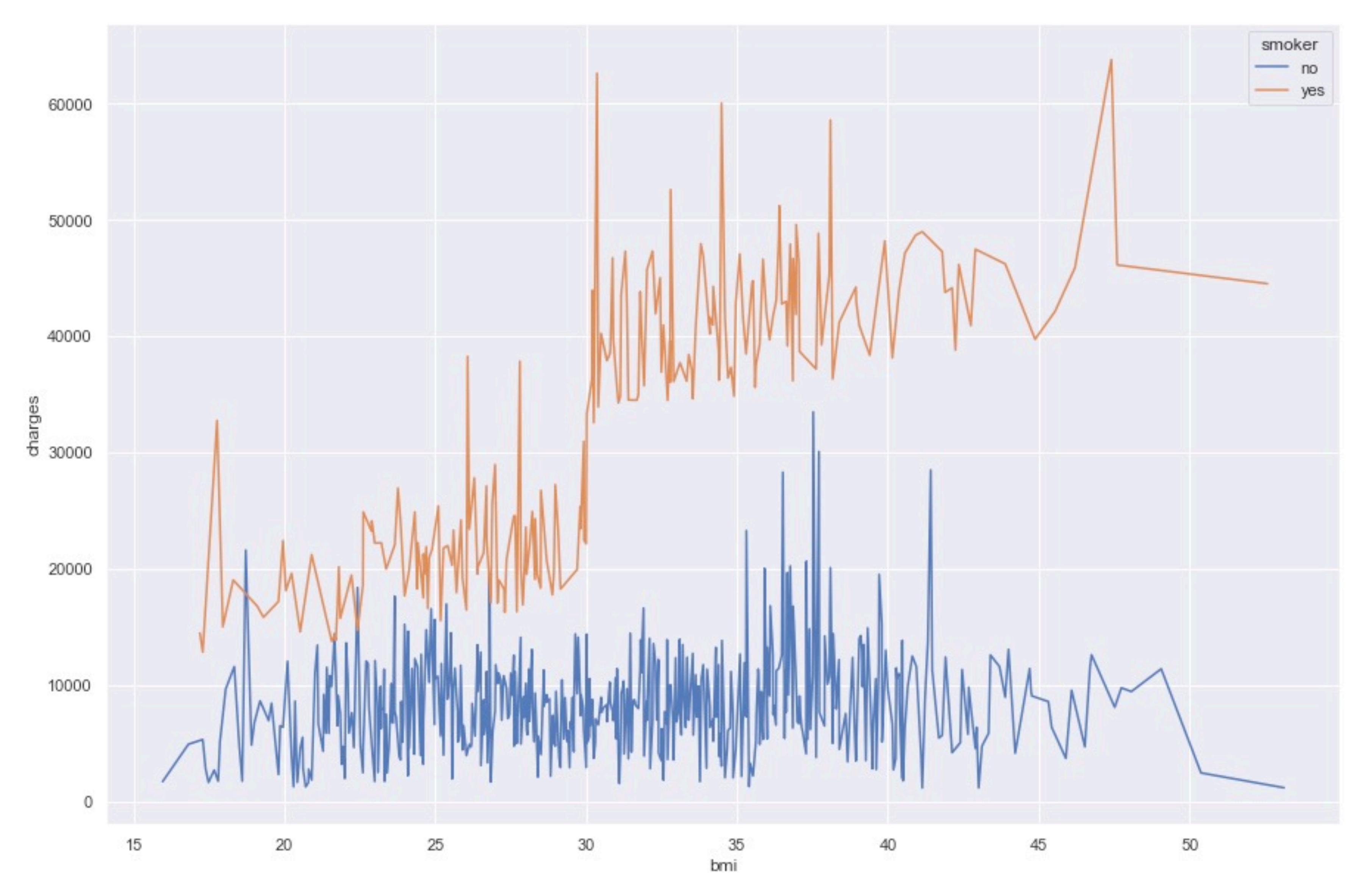
We notice two insights: (1) as age increases, charges tend to increase, and (2) smokers tend to have higher charges.



Multivariate Analysis (2 of 2)

BMI vs. Charges vs. Smoker

The BMI of smokers increases along with charges (significant jump at BMI = 30).



Hypothesis Tests (1 of 4)

- Hypothesis Definition:
    - Null Hypothesis (H_0): Mean of medical claims by smokers are **equal** to mean of medical claims by non-smokers.
    - Alternative Hypothesis (H_a): Mean of medical claims by smokers are **greater than** those of non-smokers.
- After performing an T-test (independent samples), a p-value of 2.9447e-103 was computed which is less than the 0.05 level of significance. The null hypothesis is rejected.
- **Conclusion:** The mean medical claims by smokers are greater than that of non-smokers.

Hypothesis Tests (2 of 4)

- Hypothesis Definition:
  - Null Hypothesis (H_0): Mean female BMIs is **equal** to mean male BMIs.
  - Alternative Hypothesis (H_a): Mean of female BMIs are **different than** mean of male BMIs.
- After performing an T-test (independent samples), a p-value of 0.0899 was computed which is greater than the 0.05 level of significance. The null hypothesis cannot be rejected.
- **Conclusion:** The mean female BMIs do not differ from the mean male BMIs.

Hypothesis Tests (3 of 4)

- Hypothesis Definition:
  - Null Hypothesis (H_0): Smoking and region are independent.
  - Alternative Hypothesis (H_a): Smoking and region are not independent.
- After performing a Chi-square test for independence, a p-value of 0.0617 was computed which is greater than the 0.05 level of significance. The null hypothesis cannot be rejected.
- **Conclusion:** The proportion of smokers does not differ across regions (smoking and region are independent).

Hypothesis Tests (4 of 4)

- Hypothesis Definition:
  - Null Hypothesis (H_0): The mean BMIs of women with no children, 1 child, and 2 children are **equal**
  - Alternative Hypothesis (H_a): At least one of the following does not equal the others: mean BMI of women with no children, mean BMI of women with 1 child, mean BMI of women with 2 children
- After performing a one-way ANOVA test, a p-value of 0.716 was computed which is greater than the 0.05 level of significance. The null hypothesis cannot be rejected.
- **Conclusion:** The BMI does not differ in women according to the number of children they have (means of women with no children, 1 child, and 2 children are equal).

Conclusions – Univariate Insights

- Age
  - Age ranges from 18-64, with mean and median very close in value (mean = 39.207, median = 39.000). This indicates near zero skewness.
  - There are no outliers.
  - Q3 is 51 which means that 75% of customers are below age 51.
- BMI
  - BMI ranges from 15.96-53.13, with mean and median very close in value (mean = 30.663, median = 30.400). This indicates near zero skewness.
  - There are outliers for this variable.
  - Q3 is 35 which means that 75% of customers have a BMI below 35.
- Number of Children
  - Number of children ranges from 0-5 with mean and median very close in value (mean = 1.095, median = 1.000). This indicates near zero skewness.
  - There are no outliers.
  - Q3 is 2 which means that 75% of policy holders have less than 2 children.
- Charges
  - Charges range from $1,121.87-$63,770.43, a wide range. With the mean ($13,270.42) greater than the median ($9,382.03), the data is right-skewed.
  - There are outliers for this variable all on the higher end (above $35,000).
  - The mean is greater than the median, therefore the data is right-skewed.
  - Q3 is $16,639.91. 75% of customer have been charged less than the value of Q3.
- Sex
  - More males are policy holders (676) compared to females (662). There is a 1% difference (49.5% vs. 50.5%).
- Smoker
  - Most policy holders do not smoke (1064). Non-smokers greatly outnumber smokers (79.5% of policy holders do not smoke).
- Region
  - Most policy holders live in the southeast region of the U.S. (364). The remainder of the policy holders are evenly split between the remaining regions.

Conclusions – Insights from Multivariate Analysis and Hypothesis Testing

- As age increases, charges tend to increase.
- Smokers tend to have higher charges. Based on hypothesis testing we can say that the medical claims of smokers are greater than those of non-smokers.
- Based on hypothesis testing we know that the proportion of smokers does not differ across the various regions.
- The BMI of smokers increases along with charges (significant jump at BMI = 30).
- Based on hypothesis testing we know that BMIs do not significantly differ based on gender.
- Based on hypothesis testing we know that BMI of women does not differ according to the number of children the policy holder has.

Recommendations

- Because there are more customers in the southeast region, determine if there are effective marketing campaigns or strategies unique to this region that can be utilized for other regions.
- Even though the vast majority of policy holders are non-smokers, smokers tend to have higher charges/claims than non-smokers. Consider offering special rates to smokers who are members of a program to quit smoking (potentially partner with such a program). This may also be attractive to potential policy holders.
- Charges/claims tend to increase with age. Consider researching programs that meet the needs of older policy holders to offer as incentive to maintain health.