Национальный исследовательский университет

«Высшая школа экономики»

Московский институт электроники и математики им. А.Н. Тихонова, Москва

Направление 01.03.04. Прикладная математика Бакалаврская программа "Прикладная математика"

Отчет о самостоятельной работе по дисциплине «Статистический анализ в Python»

Батурин Е. А

Москва 2021

Оглавление

[1. Общая постановка задачи 2](#_bookmark0)

1. [Предварительный анализ собранных данных](#_bookmark1) 2
   1. [Анализ особенностей данных: потенциальные ошибки и пропущенные значения, группы и](#_bookmark2) [выбросы](#_bookmark2) 2
      1. [Анализ количественных переменных](#_bookmark3) 2

[2.1.2. Анализ качественных переменных.](#_bookmark4) 4

* 1. [Анализ статистической связи](#_bookmark5) ..5
     1. [Графический анализ пары «числовая зависимая переменная – качественная](#_bookmark6) [независимая переменная»](#_bookmark6) ..5
     2. [Графический анализ пары «числовая зависимая переменная – числовая независимая](#_bookmark7) [переменная»](#_bookmark7) ..6
     3. [Анализ наличия корреляции между независимыми переменными](#_bookmark8) ..7

## 1. Общая постановка задачи

*Описание переменных:*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ***№*** | ***Характеристика***  ***объекта/явления*** | ***Название переменной*** | ***Шкала***  ***измерения*** | ***Роль*** |
| ***1.*** | *Расход топлива*  *(миль/галлон)* | *mpg* | *Относительная* | *Зависимая* |
| ***2.*** | *Масса(фунты)* | *weight* | *Относительная* | *независимая* |
| ***3.*** | *Разгон*  *(секунд до 60 миль/час)* | *acceleration* | *Относительная* | *независимая* |
| ***4.*** | *Количество цилиндров* | *cylinders* | *Номинальная* | *Независимая* |
| ***5.*** | *Код страны (1-США,2-Европа, 3-Япония)* | *origin* | *Номинальная* | *Независимая* |

Исследуемые данные (‘auto-mpg.csv’) были получены из набора данных, взято с сайта: <https://archive.ics.uci.edu/ml/datasets/Auto+MPG>

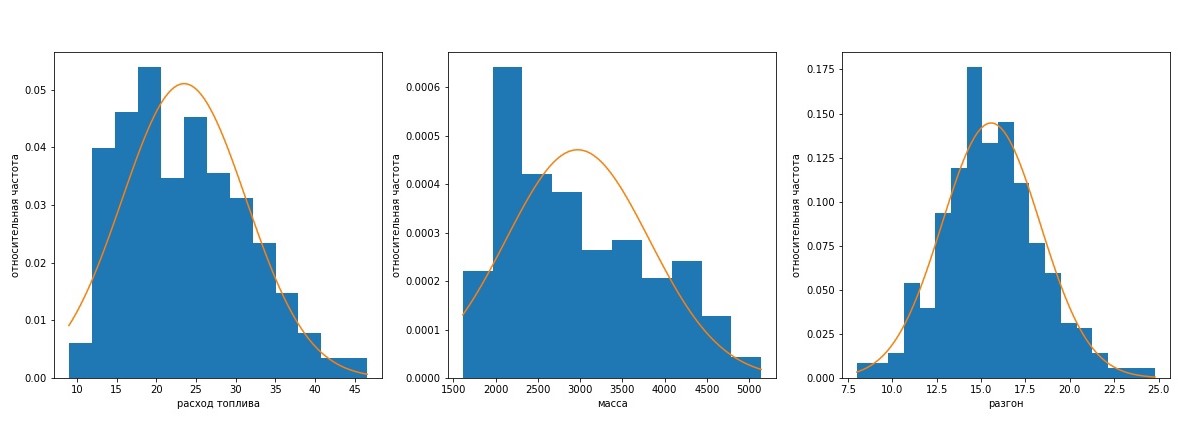
## Предварительный анализ собранных данных

* 1. Анализ особенностей данных: потенциальные ошибки и пропущенные значения, группы и выброс

В исследуемый набор данных входят 3 количественные переменные: расход топлива (mpg – зависимая переменная), масса (weight), разгон (acceleration). А также 2 качественные переменные: количество цилиндров (cylinders), код страны (origin). В данных переменных нет дубликатов и пропусков, поэтому можно приступать к их непосредственному исследованию.

* + 1. Анализ количественных переменных

Зависимая переменная «mpg» представляет собой количество миль, которые машина может проехать на одном галлоне топлива. Независимая переменная «weight» представляет собой массу автомобиля в фунтах. Последняя независимая переменная «acceleration» представляет собой количество секунд, необходимых чтобы разогнаться до скорости 60 миль/час. Далее будут представлены гистограммы распределений этих переменных.



*Рисунок 1. Гистограмма распределения количественных переменных*

|  |  |
| --- | --- |
| Статистика | Значение |
| Среднее | **15.56** |
| Медиана | **15.5** |
| Станд. отклонение | **2.75** |
| Нижний квартиль | **13.82** |
| Верхний квартиль | **17.17** |
| Межкварт. размах | **3.34** |
| Коэфф. асимметрии | **0.27** |
| Коэфф. эксцесса | **0.41** |
| Кол-во наблюдений | **398** |
| Кол-во  пропущенных значений | **0** |

|  |  |
| --- | --- |
| Статистика | Значение |
| Среднее | **2970.42** |
| Медиана | **2803.5** |
| Станд. отклонение | **846.84** |
| Нижний квартиль | **2223.75** |
| Верхний квартиль | **3608** |
| Межкварт. размах | **1384.25** |
| Коэфф. асимметрии | **0.53** |
| Коэфф. эксцесса | **-0.78** |
| Кол-во наблюдений | **398** |
| Кол-во  пропущенных значений | **0** |

|  |  |
| --- | --- |
| Статистика | Значение |
| Среднее | **23.51** |
| Медиана | **23** |
| Станд. отклонение | **7.81** |
| Нижний квартиль | **17.5** |
| Верхний квартиль | **29** |
| Межкварт. размах | **11.5** |
| Коэфф. асимметрии | **0.45** |
| Коэфф. эксцесса | **-0.51** |
| Кол-во наблюдений | **398** |
| Кол-во  пропущенных значений | **0** |

*Таблица 1. Описательная статистика количественных переменных*

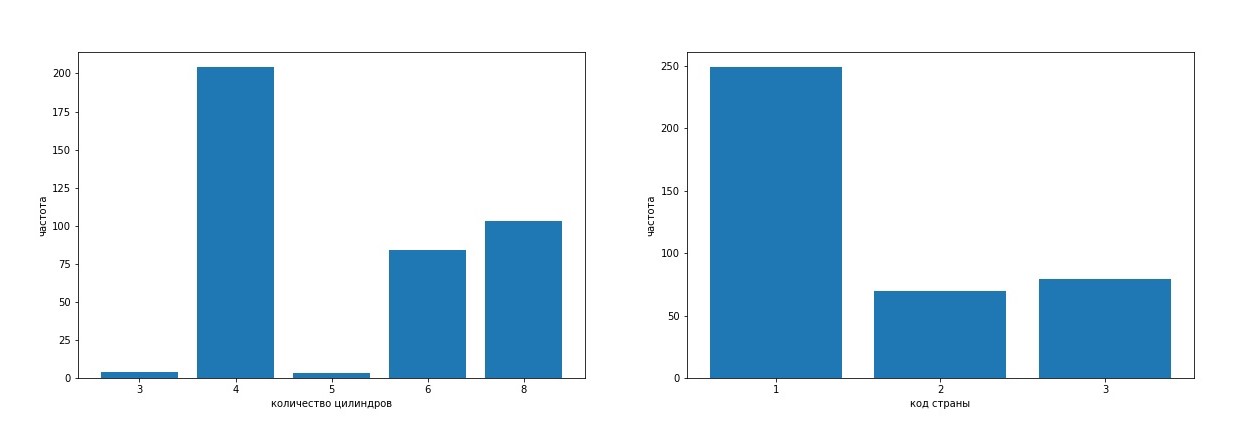
На первом графике рисунка 1 заметна полимодальность распределения (действительно, существует огромное число машин с совершенно разным расходом топлива в зависимости от огромного числа параметров), небольшая асимметрия вправо (коэффициент асимметрии равен 0.45). Второй график рисунка 1 отличается чуть большей асимметрией вправо (коэффициент асимметрии равен 0.53). Действительно, известно, что средний вес легковых автомобилей колеблется от 1.5 до 2.5 тонн. Из третьего графика видно, что распределение переменной разгона хорошо повторяет график нормального распределение со средним разгоном 15.5.

Согласно гистограммам, выбросы в данных отсутствуют. Однако, необходимо проверить это по правилу трех сигм. Для первой гистограммы имеем: 23.51+3\*7.81=46.94. В наших данных нет значений, больше 46.94, следовательно, считаем, что выбросов нет. Для второй гистограммы имеем: 2970.42+3\*846.84=5510.94. В наших данных нет значений, больше 5510.94, следовательно, считаем, что выбросов нет. Для третьей гистограммы имеем: 15.56+3\*2.75=23.81. В наших данных есть незначительное число значений, больше 23.81. Будем считать их незначительными выбросами.

Проанализировав данные на нормальность, делаем вывод, что что гипотеза о согласованности данных с нормальным распределением опровергается во всех трех случаях, так как полученные значения уровня значимости меньше 0.05(p1=1.18e-07, p2=1.97e-11, p3=0.03(хотя третье распределение хорошо повторяет нормальное)). Сравниваем с 0.05, потому что это значение является наиболее распространённым значением уровня значимости для принятия решения об опровержении гипотезы

2.1.2. Анализ качественных переменных.

Переменная «cylinders» определяет количество цилиндров в двигателе автомобиля. Переменная «origin» определяет страну, в которой был произведен автомобиль (*1-США,2-Европа, 3-Япония*).



*Рисунок 2. Гистограмма распределения качественных переменных*

Из первой гистограммы мы замечаем несбалансированность «классов». Это можно объяснить тем, большинство двигателей производят с четным числом цилиндров, потому что работу в них легче синхронизировать. Вторая гистограмма имеет такую же характеристику-несбалансированность, но менее выраженную. Очевидно, что так как данные собирались в США, то там можно встретить больше отечественных автомобилей.

2.1.3. Анализ репрезентативности выборки

1)расход топлива(mpg). Значение этой переменной варьируются от 10 до 45 «миль/галлон». Придется перевести в привычные нам «литры/100 км». Всем известно, что встречаются автомобили с расходом примерно 5-25 «л/100 км». Проведя вычисления (1 миля-1.6 км, 1 галлон-3.8 л), получаем диапазон примерно 4-27 «л/100 км». Это полностью соответствует реальности.

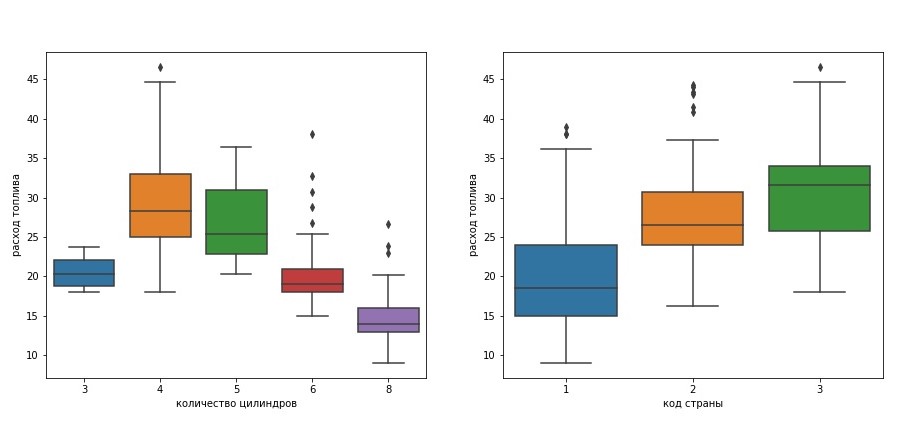
2)масса(weight). Описано в пункте 2.1.1

3)разгон(acceleration). Значение этой переменной варьируются от 7.5 до 25 «секунд до 60 миль/час». Придется перевести в привычные нам «секунд до 100 км/ч». Всем известно, что встречаются автомобили с разгоном от 5 до 25 секунд до 100 км/ч. 60 миль/ч это примерно и есть 100 км/ч, поэтому ничего переводить не нужно. Данные выборки полностью соответствует реальности.

4)кол-во цилиндров(cylinders) – описано в пункте 2.1.2

5)страна(origin)-описано в пункте 2.1.2

* 1. Анализ статистической связи
     1. Графический анализ пары «числовая зависимая переменная – качественная независимая переменная».

****

*Рисунок 3. Категорированная диаграмма Бокса-Уискера для пар расход топлива - кол-во цилиндров и расход топлива - код страны*

Из первой диаграммы видно, что в целом с увеличением количества цилиндров количество миль, которые может проехать автомобиль на одном галлоне топлива уменьшается. Это логично, ведь производители делают все более и более мощные автомобили с большим кол-вом цилиндров, которые потребляют все больше и больше топлива. Подмножество автомобилей с тремя цилиндрами не вписывается в тренд, скорее всего из-за очень малой выборки. Проверив наличие статистической связи с помощью непараметрического дисперсионного анализа с использованием критерия Краскела-Уоллиса, получаем значение p-value = 2.28е-134. Это говорит о том, что гипотеза об отсутствии связи может быть отвергнута (в качестве критического p-значения взято 5%). Таким образом, делаем вывод, что количество цилиндров влияет на количество миль, которые может проехать автомобиль на одном галлоне топлива.

Из второй диаграммы видно, что японские автомобили имеют меньший расход, чем американские и европейские. Чтобы объяснить эту зависимость, нужно быть профессионалом, разбирающимся в глубоких технических аспектах автомобилей, каковым я не являюсь, к сожалению. p-value=5.05е-136 в данном случае. Делаем аналогичный вывод, что от страны-производителя в среднем зависит расход автомобилей.

# Графический анализ пары «числовая зависимая переменная – числовая независимая переменная».

# 

*Рисунок 4. Диаграмма рассеяния для пар расход топлива – масса и расход топлива - разгон.*

Из первой диаграммы видна практически линейная зависимость расхода топлива от массы. Очевидно, чем тяжелее автомобиль, тем мощнее должен быть двигатель и тем меньше расстояние, которое может проехать автомобиль на единице топлива. Для формальной проверки гипотезы о наличии связи подсчитаем коэффициенты корреляции Пирсона и Спирмена, а также тау Кендалла и приведём результаты проверки их значимости.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Коэффициент корреляции Пирсона** | **Коэффициент корреляции Спирмена** | **Коэффициент корреляции тау Кендалла** |
| **Полученное значение** | -0.83 | -0.87 | -0.69 |
| **Значимость** | 2.97е-103 | 8.64е-127 | 2.61е-93 |

*Таблица 2. Числовой анализ связи расхода топлива и массы.*

Можно сделать вывод, что гипотеза об отсутствии статистической связи может быть отвергнута. Мы получили довольно значимые коэффициенты корреляции. Таким образом, масса автомобиля влияет на кол-во миль, которые может проехать автомобиль на единице топлива.

Из второй диаграммы видна менее четкая зависимость расхода топлива от времени разгона, но все же она есть. Чем быстрее автомобиль разгоняется, тем больше у него расход топлива. Это соответствует действительности. Посчитаем коэффициенты корреляции.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Коэффициент корреляции Пирсона** | **Коэффициент корреляции**  **Спирмена** | **Коэффициент корреляции тау**  **Кендалла** |
| **Полученное значение** | 0,42 | 0,43 | 0,3 |
| **Значимость** | 1.82е-18 | 3.77е-20 | 1.3е-18 |

*Таблица 3. Числовой анализ связи расхода топлива и разгона.*

Сделаем такой же вывод, как и из предыдущей таблицы. Параметр разгона действительно влияет на расход топлива.

* + 1. Анализ наличия корреляции между независимыми переменными.

*Связь между качественными независимыми переменными*

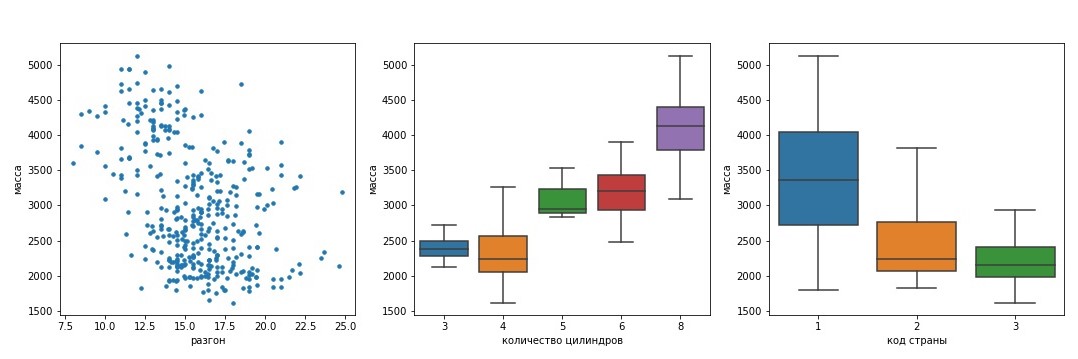
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **1** | **2** | **3** | **Total** |
| **3** | 0 | 0 | 4 | 4 |
| **4** | 72 | 63 | 69 | 204 |
| **5** | 0 | 3 | 0 | 3 |
| **6** | 74 | 4 | 6 | 84 |
| **8** | 103 | 0 | 0 | 103 |
| **Total** | 249 | 70 | 79 | 398 |

*Таблица 4. Таблица сопряженности между количеством цилиндров(строки) и страной-производителем(столбцы).*

Хи-квадрат = 180.1 (8 ст. свободы, p-значение = 9.8е-35)

Исходя из критерия хи-квадрат статистическая связь есть (p-значение < 5%). Компании из разных стран производят двигатели разных типов.

*Связь между одной из количественных переменных и всеми остальными:*



*Рисунок 5. Зависимость массы от разгона, количества цилиндров и страны.*

Из первой диаграммы видна слабая зависимость массы автомобиля от разгона. Хорошие показатели разгона могут показывать как и тяжелые автомобили с мощными двигателями, так и легкие с менее мощными двигателями. Посмотрим на коэффициенты корреляции.

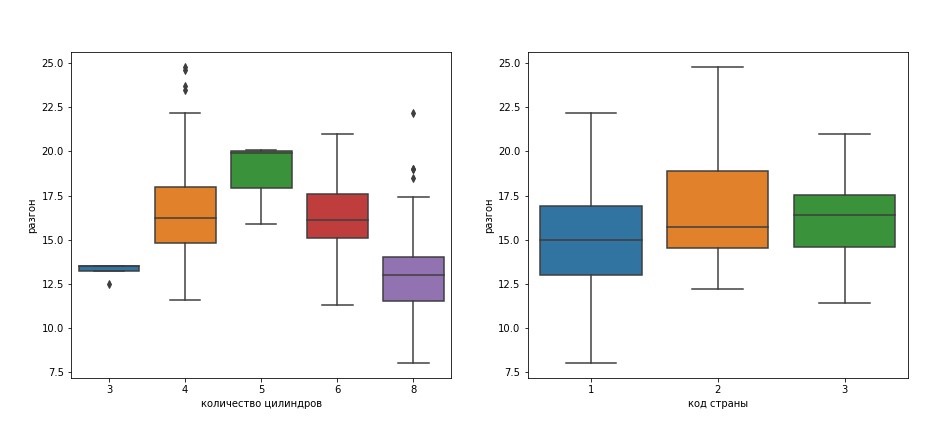
|  |  |  |  |
| --- | --- | --- | --- |
|  | **Коэффициент корреляции Пирсона** | **Коэффициент корреляции**  **Спирмена** | **Коэффициент корреляции тау**  **Кендалла** |
| **Полученное значение** | -0,41 | -0,4 | 0,26 |
| **Значимость** | 3.24е-18 | 4.18е-17 | 2.3е-15 |

Можно сделать вывод, что гипотеза об отсутствии статистической связи может быть отвергнута. Мы получили средние коэффициенты корреляции. Таким образом, скорость разгона влияет на массу автомобиля.

Из второй диаграммы видна четкая зависимость массы от кол-ва цилиндров. Действительно, чем сложнее строение двигателя, тем он мощнее, и больше масса автомобиля (в рассмотрение не берутся гоночные машины). Критерий Краскела-Уоллиса также говорит о наличии связи (p-value = 2.33е-134, что гораздо меньше порога значимости в 5%).

Из третьей диаграммы видно, что самые тяжелые машины – американские. Это явление тоже сможет объяснить только знаток автомобильной тематики. Критерий Краскела-Уоллиса также говорит о наличии связи между массой и страной (p-value = 5.18е-136, что гораздо меньше порога значимости в 5%).

*Связь между второй количественной переменной и двумя качественными:*



*Рисунок 6. Зависимости разгона от кол-ва цилиндров и страны.*

Из первой диаграммы видно, что связь присутствует. В целом, чем больше цилиндров в двигателе, тем быстрее разгон. Из тренда выбиваются автомобили с 3 и 5 цилиндрами, скорее всего, по причине малости классов. Критерий Краскела-Уоллиса также говорит о наличии связи между разгоном и количеством цилиндров (p-value = 3.3е-134, что гораздо меньше порога значимости в 5%).

Из второй диаграммы, кажется, что «родина» автомобиля практически не влияет на скорость разгона. Известно, что американские, европейские и японские компании делают одинаково хорошие автомобили. Хотя критерий Краскела-Уоллиса также говорит о наличии связи между скоростью разгона и страной-производителем (p-value = 5.02е-136, что гораздо меньше порога значимости в 5%).

Ссылка на файл с кодом: <https://drive.google.com/drive/folders/1CCbygoKI_XFAT3HScFd3kM39lYojEH2A?usp=sharing>

import matplotlib.pyplot as plt  
import numpy as np  
import pandas as pd  
import seaborn as sns  
from scipy.stats.stats import pearsonr, spearmanr, kendalltau, kruskal  
from scipy.stats import chi2\_contingency, shapiro, anderson  
  
  
*# In[39]:*data = pd.read\_csv(**'auto-mpg.csv'**)  
qualitive\_variables = [**'cylinders'**, **'origin'**]  
quantitative = data.drop(qualitive\_variables, axis=1)  
quantitive\_variables = list(quantitative)  
  
  
*# In[40]:*qualitive\_variables, quantitive\_variables  
  
  
*# In[41]:*data.head()  
  
  
*# In[42]:*len(data)  
  
  
*# In[43]:*data.info()  
  
  
*# In[44]:*data.describe()  
  
  
*# In[45]:*data.isnull().sum()  
  
  
*# In[46]:*data.median()  
  
  
*# In[47]:  
  
  
#Анализ количественных переменных*for column in quantitive\_variables:  
 quartiles = np.quantile(data[column], [0.25, 0.75])  
 print(column)  
 print(**'Нижний квартиль: '**, quartiles[0])  
 print(**'Верхний квартиль: '**, quartiles[1])  
 print(**'Межквартильный размах: '**, quartiles[1] - quartiles[0])  
 print(**'Коэффициент асимметрии: '**, data[column].skew())  
 print(**'Коэффициент эксцесса: '**, data[column].kurtosis())  
  
  
*# In[48]:  
  
  
#Гистограммы распределения количественных переменных*fig, axs = plt.subplots(nrows=1, ncols=3, figsize=(20, 6))  
labels = [**'расход топлива'**, **'масса'**, **'разгон'**]  
  
for i in range(3):  
 ax = axs[i]  
 mu = data[quantitive\_variables[i]].mean()  
 sigma = data[quantitive\_variables[i]].std()  
 x = data[quantitive\_variables[i]]  
 n, bins, patches = ax.hist(x, bins=**'sturges'**, density=True)  
 t = np.linspace(bins[0], bins[-1], 100)  
 y = ((1 / (np.sqrt(2 \* np.pi) \* sigma)) \*  
 np.exp(-0.5 \* (1 / sigma \* (t - mu))\*\*2))  
 ax.plot(t, y)  
 ax.set\_xlabel(labels[i])  
 ax.set\_ylabel(**'относительная частота'**)  
  
plt.savefig(**'histograms\_quantitive'**)  
  
  
*# In[49]:*for i in range(3):  
 \_, p = shapiro(data[quantitive\_variables[i]])  
 print(**f"p-value =** {p} **на нормальность для** {quantitive\_variables[i]}**"**)  
  
  
*# In[50]:  
  
  
#Гистограммы распределения качественных переменных*fig, axs = plt.subplots(nrows=1, ncols=2, figsize=(20, 6))  
  
values = []  
bars = list(map(str, sorted(data[**'cylinders'**].unique())))  
for j in bars:  
 values.append(len(data[data[**'cylinders'**] == int(j)]))  
ax = axs[0]  
ax.bar(bars, values)  
ax.set\_xlabel(**'количество цилиндров'**)  
ax.set\_ylabel(**'частота'**)  
  
values = []  
bars = list(map(str, sorted(data[**'origin'**].unique())))  
for j in bars:  
 values.append(len(data[data[**'origin'**] == int(j)]))  
ax = axs[1]  
ax.bar(bars, values)  
x = np.arange(len(bars))  
ax.set\_xticks(x)  
ax.set\_xticklabels(bars)  
ax.set\_xlabel(**'код страны'**)  
ax.set\_ylabel(**'частота'**)  
  
plt.savefig(**'histograms\_qualitive'**)  
  
  
*# In[51]:  
  
  
#Категорированная диаграмма Бокса-Уискера для пар расход топлива - кол-во цилиндров и расход топлива - код страны*fig, axs = plt.subplots(nrows=1, ncols=2, figsize=(15, 6))  
labels = [**'количество цилиндров'**, **'код страны'**]  
for i in range(2):  
 sns.boxplot(ax=axs[i], x=qualitive\_variables[i], y=**'mpg'**, data=data)  
 axs[i].set\_xlabel(labels[i])  
 axs[i].set\_ylabel(**'расход топлива'**)  
plt.savefig(**'boxplots'**)  
  
  
*# In[52]:*for x in qualitive\_variables:  
 \_, pvalue = kruskal(data[x], data[**'mpg'**])  
 print(**f'C использованием критерия Краскела-Уоллиса для mpg и** {x} **p-value =** {pvalue}**'**)  
  
  
*# In[53]:  
  
  
#Диаграмма рассеяния для пар расход топлива – масса и расход топлива - разгон.*fig, axs = plt.subplots(nrows=1, ncols=2, figsize=(15, 6))  
axs[0].scatter(data[**'weight'**], data[**'mpg'**], 12)  
axs[0].set\_xlabel(**'масса'**)  
axs[0].set\_ylabel(**'расход топлива'**)  
axs[1].scatter(data[**'acceleration'**], data[**'mpg'**], 12)  
axs[1].set\_xlabel(**'разгон'**)  
axs[1].set\_ylabel(**'расход топлива'**)  
plt.savefig(**'scatters'**)  
  
  
*# In[54]:*for x in [**'weight'**, **'acceleration'**]:  
 print(**f'Коэффициенты корреляции и значимость для** {x} **и mpg:'**)  
 corr, pvalue = pearsonr(data[x], data[**'mpg'**])  
 print(**f'Коэффициент корреляции Пирсона =** {corr}**, p-value =** {pvalue}**'**)  
 corr, pvalue = spearmanr(data[x], data[**'mpg'**])  
 print(**f'Коэффициент корреляции Спирмена =** {corr}**, p-value =** {pvalue}**'**)  
 corr, pvalue = kendalltau(data[x], data[**'mpg'**])  
 print(**f'Коэффициент корреляции тау Кендалла =** {corr}**, p-value =** {pvalue}**'**)  
  
  
*# In[62]:  
  
  
#Таблица сопряженности между количеством цилиндров(строки) и страной-производителем(столбцы).*print(pd.crosstab(data[**'cylinders'**], data[**'origin'**]))  
  
  
*# In[56]:*observation = pd.crosstab(data[**'cylinders'**], data[**'origin'**]).to\_numpy()  
observation  
  
  
*# In[57]:*chi2, p, dof, expected = chi2\_contingency(observation)  
print(**f'Хи-квадрат =** {chi2}**, p-value =** {p}**, степень свободы =** {dof}**'**)  
  
  
*# In[58]:  
  
  
#Зависимость массы от разгона, количества цилиндров и страны.*fig, axs = plt.subplots(nrows=1, ncols=3, figsize=(20, 6))  
axs[0].scatter(data[**'acceleration'**], data[**'weight'**], 12)  
axs[0].set\_xlabel(**'разгон'**)  
axs[0].set\_ylabel(**'масса'**)  
sns.boxplot(ax=axs[1], x=qualitive\_variables[0], y=**'weight'**, data=data)  
axs[1].set\_xlabel(**'количество цилиндров'**)  
axs[1].set\_ylabel(**'масса'**)  
sns.boxplot(ax=axs[2], x=qualitive\_variables[1], y=**'weight'**, data=data)  
axs[2].set\_xlabel(**'код страны'**)  
axs[2].set\_ylabel(**'масса'**)  
plt.savefig(**'независимые2'**)  
  
  
*# In[59]:*print(**f'Коэффициенты корреляции и значимость для acceleration и weight:'**)  
corr, pvalue = pearsonr(data[**'acceleration'**], data[**'weight'**])  
print(**f'Коэффициент корреляции Пирсона =** {corr}**, p-value =** {pvalue}**'**)  
corr, pvalue = spearmanr(data[**'acceleration'**], data[**'weight'**])  
print(**f'Коэффициент корреляции Спирмена =** {corr}**, p-value =** {pvalue}**'**)  
corr, pvalue = kendalltau(data[**'acceleration'**], data[**'weight'**])  
print(**f'Коэффициент корреляции тау Кендалла =** {corr}**, p-value =** {pvalue}**'**)  
  
  
*# In[63]:  
  
  
#Зависимости разгона от кол-ва цилиндров и страны.*fig, axs = plt.subplots(nrows=1, ncols=2, figsize=(15, 6))  
sns.boxplot(ax=axs[0], x=qualitive\_variables[0], y=**'acceleration'**, data=data)  
axs[0].set\_xlabel(**'количество цилиндров'**)  
axs[0].set\_ylabel(**'разгон'**)  
sns.boxplot(ax=axs[1], x=qualitive\_variables[1], y=**'acceleration'**, data=data)  
axs[1].set\_xlabel(**'код страны'**)  
axs[1].set\_ylabel(**'разгон'**)  
plt.savefig(**'независимые1'**)  
  
  
*# In[61]:*for x in qualitive\_variables:  
 \_, pvalue = kruskal(data[x], data[**'acceleration'**])  
 print(**f'C использованием критерия Краскела-Уоллиса для acceleration и** {x} **p-value =** {pvalue}**'**)