Министерство науки и высшего образования Российской Федерации
Калужский филиал
федерального государственного бюджетного образовательного
учреждения высшего образования
**«Московский государственный технический университет
имени Н.Э. Баумана
(национальный исследовательский университет)»**
(КФ МГТУ им. Н.Э. Баумана)

И.И. Ерохин

ФАКТОРНЫЙ АНАЛИЗ ДАННЫХ. КОРРЕЛЯЦИЯ.

Методические указания к выполнению лабораторной работы

по курсу «Технологии анализа данных»

Калуга – 2020

УДК 004.62
ББК 32.972.1
Б435

Методические указания составлены в соответствии с учебным планом КФ МГТУ им. Н.Э. Баумана по направлению подготовки 09.03.04 «Программная инженерия» кафедры «Программного обеспечения ЭВМ, информационных технологий».

Методические указания рассмотрены и одобрены:

- Кафедрой «Программного обеспечения ЭВМ, информационных технологий» (ИУ4-КФ) протокол № _ от «_ » _____ 2019 г.

Зав. кафедрой ИУ4-КФ _____к.т.н., доцент Ю.Е. Гагарин

- Методической комиссией факультета ИУ-КФ протокол №___ от «___» _____ 2020 г.

Председатель методической
комиссии факультета ИУ-КФ _____к.т.н., доцент М.Ю. Адкин

- Методической комиссией

КФ МГТУ им.Н.Э. Баумана протокол №___ от «___» _____ 2020 г.

Председатель методической
комиссии КФ МГТУ им.Н.Э.
Баумана _____д.э.н., профессор О.Л. Перерва

Рецензент:

к.т.н., доцент кафедры ИУ3-КФ _____ А.В. Финошин

Авторы

ассистент кафедры ИУ4-КФ _____ И.И. Ерохин

Аннотация

Методические указания к выполнению лабораторной работы по курсу «Технологии анализа данных» содержат общие сведения о факторном анализе и средствах языка Python для его выполнения.

Предназначены для студентов 4-го курса бакалавриата КФ МГТУ им. Н.Э. Баумана, обучающихся по направлению подготовки 09.03.04 «Программная инженерия».

© Калужский филиал МГТУ им. Н.Э. Баумана, 2020 г.

© И.И. Ерохин, 2020 г.

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Настоящие методические указания составлены в соответствии с программой проведения лабораторных работ по курсу «Технологии анализа данных» на кафедре «Программное обеспечение ЭВМ, информационные технологии» факультета «Информатика и управление» Калужского филиала МГТУ им. Н.Э. Баумана.

Методические указания, ориентированные на студентов 4-го курса направления подготовки 09.03.04 «Программная инженерия», содержат базовые сведения о факторном анализе.

Методические указания составлены для ознакомления студентов с библиотеками, применяемые для анализа данных в языке Python. Для выполнения лабораторной работы студенту необходимы знания языка программирования Python и навыки работы с Anaconda.

ЦЕЛЬ И ЗАДАЧИ РАБОТЫ, ТРЕБОВАНИЯ К РЕЗУЛЬТАТАМ ЕЕ ВЫПОЛНЕНИЯ

Целью выполнения лабораторной работы является формирование практических навыков проведения факторного анализа и обнаружения корреляции между параметрами.

Основными задачами выполнения лабораторной работы являются:

1. Ознакомиться с понятием факторный анализ и корреляция.
2. Изучить средства языка Python для выполнения факторного анализа.

Результатами работы являются:

1. Проведенный факторный анализ.
2. Построенные графики.
3. Подготовленный отчет.

КРАТКАЯ ХАРАКТЕРИСТИКА ОБЪЕКТА ИЗУЧЕНИЯ, ИССЛЕДОВАНИЯ

Факторный анализ — многомерный метод, применяемый для изучения взаимосвязей между значениями переменных. Предполагается, что известные переменные зависят от меньшего количества неизвестных переменных и случайной ошибки.

Факторный анализ позволяет решить две важные проблемы исследователя: описать объект измерения всесторонне и в то же время компактно. С помощью факторного анализа возможно выявление скрытых переменных факторов, отвечающих за наличие линейных статистических корреляций между наблюдаемыми переменными.

Две основных цели факторного анализа:

- определение взаимосвязей между переменными;
- сокращение числа переменных необходимых для описания данных.

Практическое выполнение факторного анализа начинается с проверки его условий. В обязательные условия факторного анализа входят:

- все признаки должны быть количественными;
- число наблюдений должно быть не менее чем в два раза больше числа переменных;
- выборка должна быть однородна;
- исходные переменные должны быть распределены симметрично;
- факторный анализ осуществляется по коррелирующим переменным.

Стандартный способ оценки воздействия факторов - регрессия методом наименьших квадратов. Для этого в пакете statsmodels имеется функция ols, которой необходимо передать формулу регрессии, а также фрейм данных.

РЕГРЕССИЯ И РЕГРЕССИОННЫЙ АНАЛИЗ

Регре́ссия (лат. regressio — обратное движение, отход) в теории вероятностей и математической статистике — математическое выражение, отражающее зависимость зависимой переменной y от независимых переменных x при условии, что это выражение будет иметь статистическую значимость. В отличие от чисто функциональной зависимости y=f(x), когда каждому значению независимой переменной x соответствует одно определённое значение величины y, при регрессионной связи одному и тому же значению x могут соответствовать в зависимости от случая различные значения величины y.

В общем случае регрессия одной случайной переменной на другую не обязательно будет линейной. Также не обязательно ограничиваться парой случайных переменных. Статистические проблемы регрессии связаны с определением общего вида уравнения регрессии, построением оценок неизвестных параметров, входящих в уравнение регрессии, и проверкой статистических гипотез о регрессии. Эти проблемы рассматриваются в рамках регрессионного анализа.

Регрессионный анализ — статистический метод исследования влияния одной или нескольких независимых переменных X_1, X_2, … X_P на зависимую переменную Y. Независимые переменные иначе называют регрессорами или предикторами, а зависимые переменные — критериальными. Терминология зависимых и независимых переменных отражает лишь математическую зависимость переменных, а не причинно-следственные отношения.

Цели регрессионного анализа:

- Определение степени детерминированности вариации критериальной (зависимой) переменной предикторами (независимыми переменными)
- Предсказание значения зависимой переменной с помощью независимой(-ых)
- Определение вклада отдельных независимых переменных в вариацию зависимой

Примеры применения регрессионного анализа:

- Моделирование потоков миграции в зависимости от таких факторов как средний уровень зарплат, наличие медицинских, школьных учреждений, географическое положение…
- Моделирование дорожных аварий как функции скорости, дорожных условий, погоды и т.д.,
- Моделирование потерь от пожаров как функции от таких переменных как количество пожарных станций, время обработки вызова, или цена собственности.

Суть регрессионного анализа заключается в нахождении наиболее важных факторов, которые влияют на зависимую переменную.

Уравнение регрессии - это математическая формула, применяемая к независимым переменным, чтобы лучше спрогнозировать зависимую переменную, которую необходимо смоделировать:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon$$

Термины и концепции регрессионного анализа

Зависимая переменная (Y) - это переменная, описывающая процесс, который мы пытаемся предсказать или понять.

Независимые переменные (X) - это переменные, используемые для моделирования или прогнозирования значений зависимых переменных. В уравнении регрессии они располагаются справа от знака равенства и часто называются объяснительными переменными. Зависимая переменная - это функция независимых переменных.

Коэффициенты регрессии (β) - это коэффициенты, которые рассчитываются в результате выполнения регрессионного анализа. Вычисляются величины для каждой независимой переменной, которые представляют силу и тип взаимосвязи независимой переменной по отношению к зависимой.

Невязки. Существует необъяснимое количество зависимых величин, представленных в уравнении регрессии как случайные ошибки **ε**.

Создание регрессионной модели представляет собой итерационный процесс, направленный на поиск эффективных независимых переменных, чтобы объяснить зависимые переменные, которые необходимо смоделировать или понять, запуская инструмент регрессии, чтобы определить, какие величины являются эффективными предсказателями. Затем пошаговое удаление и/или добавление переменных до тех пор, пока не будет найдена наилучшим образом подходящая регрессионная модель. Процесс построения регрессионной модели должен учитывать теоретические аспекты, мнение экспертов в этой области и здравый смысл.

Метод наименьших квадратов

Метод наименьших квадратов (МНК) — математический метод, применяемый для решения различных задач, основанный на минимизации суммы квадратов отклонений некоторых функций от искомых переменных. МНК является одним из базовых методов регрессионного анализа для оценки неизвестных параметров регрессионных моделей по выборочным данным. МНК позволяет получить такие оценки параметров, при которых сумма квадратов отклонений фактических значений результативного признака y от теоретических минимальна.

Пусть x — набор n неизвестных переменных (параметров), f_i (x), i=1, …, m, m> n — совокупность функций от этого набора переменных. Задача заключается в подборе таких значений x, чтобы значения этих функций были максимально близки к некоторым значениям yi. По существу, речь идет о «решении» переопределенной системы уравнений $f_i(x) = y_i$, i = 1,..., m в указанном смысле максимальной близости левой и правой частей системы. Суть МНК заключается в выборе в качестве «меры близости» суммы квадратов отклонений левых и правых частей $|f_i(x) - y_i|$. Таким образом, сущность МНК может быть выражена следующим образом:

$$\sum_i e_i^2 = \sum_i ¿¿$$

В случае, если система уравнений имеет решение, то наименьшее значение суммы квадратов будет равно нулю и могут быть найдены точные решения системы уравнений аналитически или. например, различными численными методами оптимизации. Если система переопределена, то есть, говоря нестрого, количество независимых уравнений больше количества искомых переменных, то система не имеет точного решения и метод наименьших квадратов позволяет найти некоторый «оптимальный» вектор *x* в смысле максимальной близости векторов *y* и *f(x)* или максимальной близости вектора отклонений *e* к нулю (близость понимается в смысле евклидова расстояния).

Интерпретация значений регрессионного анализа

В общем случае **коэффициент регрессии k** показывает, как в среднем изменится результативный признак Y, если факторный признак X увеличится на единицу.

Свойства коэффициента регрессии:

- Коэффициент регрессии может принимать любые значения.
- Коэффициент регрессии не симметричен, т.е. изменяется, если X и Y поменять местами.
- Единицей измерения коэффициента регрессии является отношение единицы измерения Y к единице измерения X: ([Y] / [X]).
- Коэффициент регрессии изменяется при изменении единиц измерения X и Y.

Y пересечение (intercept) – коэффициент который показывает какой будет Y в случае, если все используемые в модели факторы будут равны 0, подразумевается, что это зависимость от других неописанных
в модели факторов.

R^2 - коэффициент детерминации, показывающий на сколько процентов расчетные параметры модели, то есть сама модель, объясняют зависимость и изменения изучаемого параметра -Y от исследуемых факторов - X. Можно сказать, что, это показатель качества модели и чем он выше, тем лучше. Понятное дело, что он не может быть больше 1 и считается неплохо, когда R^2 выше 0,8, а если меньше 0,5, то смысл такой модели можно смело ставить под большой вопрос.

Коэффициент детерминации рассматривают, как правило, в качестве основного показателя, отражающего меру качества регрессионной модели, описывающей связь между зависимой и независимыми переменными модели. Коэффициент детерминации показывает, какая доля вариации объясняемой переменной y учтена в модели и обусловлена влиянием на нее факторов, включенных в модель:

$$R^2=1-\sum_{i=1}^{n} ¿¿¿$$

где - y_i– значения наблюдаемой переменной,

$\overline{y_i}$ – среднее значение по наблюдаемым данным,

$\widehat{y}_i$ – модельные значения, построенные по оцененным параметрам.

Столбец **p-значение** отражает достоверность отличия соответствующих коэффициентов от нуля. В случае, когда p>0,05, коэффициент может считаться нулевым. Это означает, что соответствующая независимая переменная практически не влияет на зависимую переменную и коэффициент может быть убран из уравнения.

Значение **Prob (F-statistic) <0.05** позволяет отвергнуть гипотезу о незначимости регрессии.

КОРРЕЛЯЦИЯ

Корреляция (от лат. correlatio «соотношение, взаимосвязь») или корреляционная зависимость — статистическая взаимосвязь двух или более случайных величин (либо величин, которые можно с некоторой допустимой степенью точности считать таковыми). При этом изменения значений одной или нескольких из этих величин сопутствуют систематическому изменению значений другой или других величин.

Математической мерой корреляции двух случайных величин служит корреляционное отношение η либо коэффициент корреляции R. В случае если изменение одной случайной величины не ведёт к закономерному изменению другой случайной величины, но приводит к изменению другой статистической характеристики данной случайной величины, то подобная связь не считается корреляционной, хотя и является статистической.

Значительная корреляция между двумя случайными величинами всегда является свидетельством существования некоторой статистической связи в данной выборке, но эта связь не обязательно должна наблюдаться для другой выборки и иметь причинно-следственный характер. Часто заманчивая простота корреляционного исследования подталкивает исследователя делать ложные интуитивные выводы о наличии причинно-следственной связи между парами признаков, в то время как коэффициенты корреляции устанавливают лишь статистические взаимосвязи. Корреляция двух величин может свидетельствовать о существовании общей причины, хотя сами явления напрямую не взаимодействуют.

В то же время, отсутствие корреляции между двумя величинами ещё не значит, что между ними нет никакой связи. Например, зависимость может иметь сложный нелинейный характер, который корреляция не выявляет.

Некоторые виды коэффициентов корреляции могут быть положительными или отрицательными. В первом случае предполагается, что можно определить только наличие или отсутствие

связи, а во втором — также и её направление. Если предполагается, что на значениях переменных задано отношение строгого порядка, то отрицательная корреляция — корреляция, при которой увеличение одной переменной связано с уменьшением другой. При этом коэффициент корреляции будет отрицательным. Положительная корреляция в таких условиях — это такая связь, при которой увеличение одной переменной связано с увеличением другой переменной. Возможна также ситуация отсутствия статистической взаимосвязи — например, для независимых случайных величин.

Корреляционная матрица

При большом числе наблюдений, когда коэффициенты корреляции необходимо последовательно вычислять из нескольких рядов числовых данных, для удобства получаемые коэффициенты сводят в таблицы, называемые корреляционными матрицами.

Корреляционная матрица — это квадратная (или прямоугольная) таблица, в которой на пересечении соответствующих строки и столбца находится коэффициент корреляции между соответствующими параметрами.

Корреляционная матрица является симметричной, с единичной главной диагональю, положительно полуопределенной матрицей.

ОБРАЗЕЦ ВЫПОЛНЕНИЯ ЗАДАНИЯ

```
import pandas_datareader.data as web

import statsmodels.formula.api as smf

import statsmodels.api as sm

from statsmodels.graphics.regressionplots import plot_fit

import matplotlib.pyplot as plt

import seaborn as sns

#факторный анализ при помощи линейной регрессии

moex = web.DataReader('USD000UTSTOM', 'moex', start='1999-01-01',
end='2018-12-31')

data = moex[['OPEN', 'LOW', 'HIGH', 'CLOSE',
'NUMTRADES']].loc[moex['BOARDID'] == 'CETS']

result = smf.ols(formula="NUMTRADES ~ OPEN + LOW + HIGH + CLOSE",
data=data).fit()

#вывод результатов регрессии

print(result.params)

print(result.params.values)

print(result.summary())

#построение столбчатой диаграммы

D = {}

y = result.params.values[1:]

for i in range(0, len(y)):

    D[data.columns[i]] = abs(y[i])

plt.bar(range(len(D)), D.values(), align='center')

plt.xticks(range(len(D)), D.keys())

corr = data.corr()

print(corr)

sns.heatmap(corr,

            xticklabels=corr.columns.values,

            yticklabels=corr.columns.values)
```

В результате выполнения программы будут получены следующие результаты (рис. 1, 2, 3):

```
Intercept   -22282.446561
OPEN          6129.518735
LOW         -12760.743701
HIGH          1750.994125
CLOSE         5773.064544
dtype: float64
[-22282.44656081   6129.51873509 -12760.74370086   1750.99412489
   5773.06454398]
                            OLS Regression Results
==============================================================================
Dep. Variable:              NUMTRADES   R-squared:                       0.796
Model:                            OLS   Adj. R-squared:                  0.796
Method:                 Least Squares   F-statistic:                     3826.
Date:                Sat, 09 Mar 2019   Prob (F-statistic):               0.00
Time:                        15:28:41   Log-Likelihood:                -41042.
No. Observations:                3925   AIC:                         8.209e+04
Df Residuals:                    3920   BIC:                         8.213e+04
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept  -2.228e+04    400.750    -55.602      0.000   -2.31e+04   -2.15e+04
OPEN        6129.5187    619.927      9.887      0.000    4914.109    7344.928
LOW        -1.276e+04    762.372    -16.738      0.000   -1.43e+04   -1.13e+04
HIGH        1750.9941    499.337      3.507      0.000     772.008    2729.980
CLOSE       5773.0645    682.640      8.457      0.000    4434.702    7111.427
==============================================================================
Omnibus:                      778.222   Durbin-Watson:                   0.524
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            16743.917
Skew:                           0.362   Prob(JB):                         0.00
Kurtosis:                      13.092   Cond. No.                         718.
==============================================================================
```

Рис. 1. Результаты регрессии

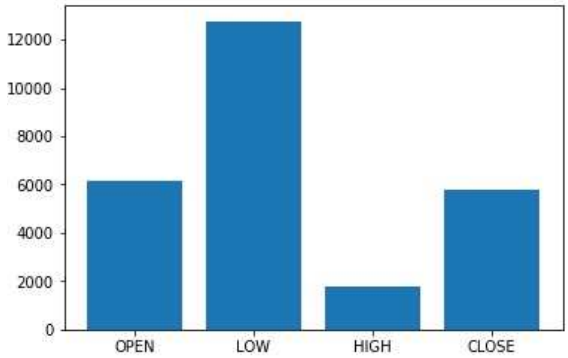

Рис. 2. Диаграмма степени влияния параметров на исследуемую величину (в абсолютном отношении)

	OPEN	LOW	HIGH	CLOSE	NUMTRADES
OPEN	1.000000	0.999775	0.999553	0.999561	0.872382
LOW	0.999775	1.000000	0.999339	0.999753	0.869567
HIGH	0.999553	0.999339	1.000000	0.999659	0.875718
CLOSE	0.999561	0.999753	0.999659	1.000000	0.872630
NUMTRADES	0.872382	0.869567	0.875718	0.872630	1.000000

```
Out[6]: <matplotlib.axes._subplots.AxesSubplot at 0x7f451da42908>
```

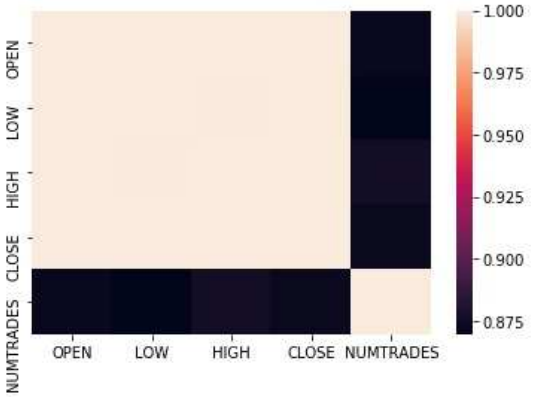

Рис. 3. Корреляционная матрица и ее визуализация

ЗАДАНИЕ НА ЛАБОРАТОРНУЮ РАБОТУ

Для всех вариантов необходимо осуществить факторный анализ данных при помощи регрессии методом наименьших квадратов. Визуализировать полученные данные в виде графиков и диаграмм. Файлы для заданий имеют названия, соответствующие варианту.

ТРЕБОВАНИЯ К РЕАЛИЗАЦИИ

В качестве результата работы необходимо построить графики (диаграммы) и корреляционную матрицу и отобразить их в Jupyter Notebook. Также необходимо проанализировать полученные результаты и сделать выводы.

ВАРИАНТЫ ЗАДАНИЙ

Вариант 1

Считать данные IEX (pandas_datareader) в структуру DataFrame. Провести факторный анализ зависимости параметра volume от 2 характеристик: high, low. Вывести результаты (summary) анализа и интерпретировать их. Построить столбчатую диаграмму, отражающую степень влияния каждого параметра на значение volume. Повторить анализ для данных разного размера (необходимо изменить значения начальной и конечной даты). Построить график зависимости коэффициента детерминации (R-square) от размера набора данных. Сделать выводы. Построить и визуализировать корреляционную матрицу для всех параметров (high, low, volume). Сделать выводы о наличии связей между параметрами.

Вариант 2

Считать данные из CSV файла в структуру DataFrame. Провести факторный анализ зависимости числа просмотров от 2 характеристик: загруженных видео и числа подписчиков. Вывести результаты (summary) анализа и интерпретировать их. Построить столбчатую диаграмму, отражающую степень влияния каждого параметра на число просмотров. Повторить анализ для данных разного размера.

Построить график зависимости коэффициента детерминации (R-square) от размера набора данных. Сделать выводы. Построить и визуализировать корреляционную матрицу для всех параметров (просмотры, подписчики, загруженные видео). Сделать выводы о наличии связей между параметрами.

Вариант 3

Считать данные из CSV файла в структуру DataFrame. Провести факторный анализ зависимости параметра Spending Score от 2 характеристик: возраста и годового дохода. Вывести результаты (summary) анализа и интерпретировать их. Построить столбчатую диаграмму, отражающую степень влияния каждого параметра на Spending Score. Повторить анализ для данных разного размера. Построить график зависимости коэффициента детерминации (R-square) от размера набора данных. Сделать выводы. Построить и визуализировать корреляционную матрицу для всех параметров (Spending Score, возраст, доход). Сделать выводы о наличии связей между параметрами.

Вариант 4

Считать данные из CSV файла в структуру DataFrame. Провести факторный анализ зависимости стоимости футболиста от 2 характеристик: возраста и общего рейтинга. Вывести результаты (summary) анализа и интерпретировать их. Построить столбчатую диаграмму, отражающую степень влияния каждого параметра на стоимость. Повторить анализ для данных разного размера. Построить график зависимости коэффициента детерминации (R-square) от размера набора данных. Сделать выводы. Построить и визуализировать корреляционную матрицу для всех параметров (стоимость, возраст, рейтинг). Сделать выводы о наличии связей между параметрами.

Вариант 5

Считать данные из CSV файла в структуру DataFrame. Провести факторный анализ зависимости параметра Purchase от 2 характеристик: Occupation и Stay_In_Current_City_Years. Вывести результаты (summary) анализа и интерпретировать их. Построить столбчатую диаграмму, отражающую степень влияния каждого параметра на число Purchase. Повторить анализ для данных разного размера. Построить график зависимости коэффициента детерминации (R-square) от размера набора данных. Сделать выводы. Построить и визуализировать корреляционную матрицу для всех параметров (Purchase, Occupation, Stay_In_Current_City_Years). Сделать выводы о наличии связей между параметрами.

Вариант 6

Считать данные из CSV файла в структуру DataFrame. Провести факторный анализ зависимости параметра writing score от 2 характеристик: reading score и math score. Вывести результаты (summary) анализа и интерпретировать их. Построить столбчатую диаграмму, отражающую степень влияния каждого параметра на число writing score. Повторить анализ для данных разного размера. Построить график зависимости коэффициента детерминации (R-square) от размера набора данных. Сделать выводы. Построить и визуализировать корреляционную матрицу для всех параметров (reading score, writing score, math score). Сделать выводы о наличии связей между параметрами.

Вариант 7

Считать данные из CSV файла в структуру DataFrame. Провести факторный анализ зависимости параметра KMs Driven от 2 характеристик: Price и Year. Вывести результаты (summary) анализа и интерпретировать их. Построить столбчатую диаграмму, отражающую степень влияния каждого параметра на число KMs Driven. Повторить анализ для данных разного размера. Построить график зависимости коэффициента детерминации (R-square) от размера набора данных. Сделать выводы. Построить и

визуализировать корреляционную матрицу для всех параметров (KMs Driven, Price, Year). Сделать выводы о наличии связей между параметрами.

Вариант 8

Считать данные IEX (pandas_datareader) в структуру DataFrame. Провести факторный анализ зависимости параметра volume от 2 характеристик: open, close. Вывести результаты (summary) анализа и интерпретировать их. Построить столбчатую диаграмму, отражающую степень влияния каждого параметра на значение volume. Повторить анализ для данных разного размера (необходимо изменить значения начальной и конечной даты). Построить график зависимости коэффициента детерминации (R-square) от размера набора данных. Сделать выводы. Построить и визуализировать корреляционную матрицу для всех параметров (open, close, volume). Сделать выводы о наличии связей между параметрами.

Вариант 9

Считать данные из CSV файла в структуру DataFrame. Провести факторный анализ зависимости числа загрузок (активности) от 2 характеристик: числа просмотров и подписчиков. Вывести результаты (summary) анализа и интерпретировать их. Построить столбчатую диаграмму, отражающую степень влияния каждого параметра на активность канала. Повторить анализ для данных разного размера. Построить график зависимости коэффициента детерминации (R-square) от размера набора данных. Сделать выводы. Построить и визуализировать корреляционную матрицу для всех параметров (просмотры, подписчики, загруженные видео). Сделать выводы о наличии связей между параметрами.

Вариант 10

Считать данные из CSV файла в структуру DataFrame. Провести факторный анализ зависимости параметра годовой доход от 2 характеристик: возраста и Spending Score. Вывести результаты

(summary) анализа и интерпретировать их. Построить столбчатую диаграмму, отражающую степень влияния каждого параметра на годовой доход. Повторить анализ для данных разного размера. Построить график зависимости коэффициента детерминации (R-square) от размера набора данных. Сделать выводы. Построить и визуализировать корреляционную матрицу для всех параметров (Spending Score, возраст, доход). Сделать выводы о наличии связей между параметрами.

Вариант 11

Считать данные из CSV файла в структуру DataFrame. Провести факторный анализ зависимости рейтинга футболиста от 2 характеристик: стоимости и зарплаты. Вывести результаты (summary) анализа и интерпретировать их. Построить столбчатую диаграмму, отражающую степень влияния каждого параметра на рейтинг. Повторить анализ для данных разного размера. Построить график зависимости коэффициента детерминации (R-square) от размера набора данных. Сделать выводы. Построить и визуализировать корреляционную матрицу для всех параметров (стоимость, возраст, зарплата). Сделать выводы о наличии связей между параметрами.

Вариант 12

Считать данные из CSV файла в структуру DataFrame. Провести факторный анализ зависимости параметра math score от 2 характеристик: reading score и writing score. Вывести результаты (summary) анализа и интерпретировать их. Построить столбчатую диаграмму, отражающую степень влияния каждого параметра на число math score. Повторить анализ для данных разного размера. Построить график зависимости коэффициента детерминации (R-square) от размера набора данных. Сделать выводы. Построить и визуализировать корреляционную матрицу для всех параметров (reading score, writing score, math score). Сделать выводы о наличии связей между параметрами.

Вариант 13

Считать данные из CSV файла в структуру DataFrame. Провести факторный анализ зависимости параметра Price от 2 характеристик: KMs Driven и Year. Вывести результаты (summary) анализа и интерпретировать их. Построить столбчатую диаграмму, отражающую степень влияния каждого параметра на число Price. Повторить анализ для данных разного размера. Построить график зависимости коэффициента детерминации (R-square) от размера набора данных. Сделать выводы. Построить и визуализировать корреляционную матрицу для всех параметров (KMs Driven, Price, Year). Сделать выводы о наличии связей между параметрами.

Вариант 14

Считать данные из CSV файла в структуру DataFrame. Провести факторный анализ зависимости зарплаты футболиста от 2 характеристик: стоимости и общего рейтинга. Вывести результаты (summary) анализа и интерпретировать их. Построить столбчатую диаграмму, отражающую степень влияния каждого параметра на зарплату. Повторить анализ для данных разного размера. Построить график зависимости коэффициента детерминации (R-square) от размера набора данных. Сделать выводы. Построить и визуализировать корреляционную матрицу для всех параметров (стоимость, зарплата, рейтинг). Сделать выводы о наличии связей между параметрами.

Вариант 15

Считать данные IEX (pandas_datareader) в структуру DataFrame. Провести факторный анализ зависимости параметра volume от 4 характеристик: open, high, low, close. Вывести результаты (summary) анализа и интерпретировать их. Построить столбчатую диаграмму, отражающую степень влияния каждого параметра на значение volume. Повторить анализ для данных разного размера (необходимо изменить значения начальной и конечной даты). Построить график зависимости коэффициента детерминации (R-square) от размера набора данных.

Сделать выводы. Построить и визуализировать корреляционную матрицу для всех параметров (open, high, low, close, volume). Сделать выводы о наличии связей между параметрами.

КОНТРОЛЬНЫЕ ВОПРОСЫ И ЗАДАНИЯ

1. Дайте определение понятию факторный анализ.
2. Дайте определение понятию регрессия.
3. Перечислите основные цели регрессионного анализа.
4. Перечислите и раскройте основные понятия регрессионного анализа.
5. Раскройте понятие уравнение регрессии.
6. Раскройте сущность метода наименьших квадратов (МНК).
7. Раскройте сущность и назначение коэффициента регрессии.
8. Раскройте сущность и назначение коэффициента детерминации.
9. Дайте определение понятию пересечение.
10. Опишите назначение параметров p-значение и Prob (F-statistic).
11. Раскройте сущность понятия корреляция.
12. Раскройте, что означает значительная корреляция между величинами.
13. Дайте определение понятию корреляционная матрица и перечислите её основные свойства.
14. Назовите функцию, которая используется в Python для проведения регрессионного анализа методом наименьших квадратов.

ФОРМА ОТЧЕТА ПО ЛАБОРАТОРНОЙ РАБОТЕ

На выполнение лабораторной работы отводится 1 занятие (2 академических часа: 1 час на выполнение и сдачу лабораторной работы и 1 час на подготовку отчета).

Отчет на защиту предоставляется в печатном виде.

Структура отчета (на отдельном листе(-ах)): титульный лист, формулировка задания, описание процесса выполнения лабораторной работы, результаты выполнения работы, выводы.

ОСНОВНАЯ ЛИТЕРАТУРА

1. Маккинли, Уэс Python и анализ данных / Пер. с англ. Слинкин А.А. - М.: ДМК Пресс, 2015. - 482 с.:ил.
2. Грас, Дж. Data Science. Наука о данных с нуля / Пер. с англ. - СПБ.: БХВ -Петербург, 2017. - 336с.: ил.

ДОПОЛНИТЕЛЬНАЯ ЛИТЕРАТУРА

3. Henley, A.J. Learn Data Analysis with Python: Lessons in Coding / A.J. Henley, Dave Wolf ISBN 978-1-4842-3486-0

Электронные ресурсы:

4. Научная электронная библиотека http://eLIBRARY.RU
5. Электронно-библиотечная система http://e.lanbook.com