
EDA

KNIME vs Python

Выполнила
Гурская Дарина, гр. 02323-ДБ

Цель | задачи

Цель

Провести обзор и сравнительный анализ инструментов Python и ППП KNIME для проведения EDA

Задачи

1. Рассмотреть возможности Python для проведения EDA
2. Рассмотреть возможности KNIME для проведения EDA
3. Выделить плюсы и минусы
4. Дать рекомендации по применению

EDA

Разведочный анализ данных
Exploratory data analysis (EDA)

Инструменты и методы

визуализация данных

сводные статистики и меры центральной
тенденции

анализ выбросов и аномалий

корреляционный анализ

Данные

В качестве данных для проведения EDA были использованы данные о посуточной аренде квартир в Иркутске

Данные взяты с портала sutochno.ru и актуальны на 19 февраля 2024 года

№	Район	Оценка	Стоимость, руб./сут.	Предоплата, руб.	Площадь, м^2	Кол-во гостей	Курение	Вечеринки	балкон / лоджия	Спальные места	Этаж	Кол-во кроватей	Питомцы	Лифт	Парковка	Страховой депозит	Ремонт
2	Октябрьский	9,40	2520	630	25	2	нет	да	да	1	Верхний	1	нет	да	да	1000	евроремонт
3	Октябрьский	10,00	2300	460	30	2	нет	нет	да	2	Верхний	1	да	да	да	1000	евроремонт
4	Октябрьский	10,00	3300	660	48	2	нет	да	да	2	Верхний	2	да	да	да	1000	евроремонт
5	Октябрьский	9,90	2290	458	42	4	нет	нет	да	2	Верхний	2	да	да	да	1000	евроремонт
6	Октябрьский	10,00	4125	1031	37	4	нет	да	да	2	Верхний	2	нет	да	да	1000	евроремонт

Python

Язык программирования, который предлагает множество инструментов для решения разнообразных задач, включая задачи статистического анализа

Основные библиотеки для работы со статистикой в Python:

NumPy	работа с числовыми массивами и матрицами в Python
Pandas	работа с табличными данными (DataFrame) и временными рядами
Scipy.stats	статистические функции для анализа данных
Seaborn	визуализация данных
Matplotlib	

Python

```
df = pd.read_csv("Statistics.csv")  
print(df.shape)  
df.head()
```

```
price = df["Стоимость, руб./сут."  
price.head()
```

Описательные статистики

✓ [7] price.mean()

0
сек.

2938.96

✓ [8] price.median()

0
сек.

2831.0

✓ [9] min = price.min()

0
сек.

✓ [10] price.max()

0
сек.

5000

✓ [11] price.std()

0
сек.

712.4387043632228

✓ [12] price.var()

0
сек.

507568.9074747475

✓ [13] price.sum()

0
сек.

293896

✓ [14] price.quantile(0.5)

0
сек.

2831.0

✓ [54] price.quantile(0.75)

0
сек.

3360.5

```
[17] df[["Стоимость, руб./сут.", "Предоплата, руб.", "Площадь, м^2", ]].describe()
```

	Стоимость, руб./сут.	Предоплата, руб.	Площадь, м^2
count	100.00	100.00	100.00
mean	2938.96	606.70	35.50
std	712.44	205.77	8.87
min	1300.00	260.00	18.00
25%	2447.50	458.00	30.00
50%	2831.00	594.00	36.00
75%	3360.50	720.00	40.25
max	5000.00	1500.00	64.00



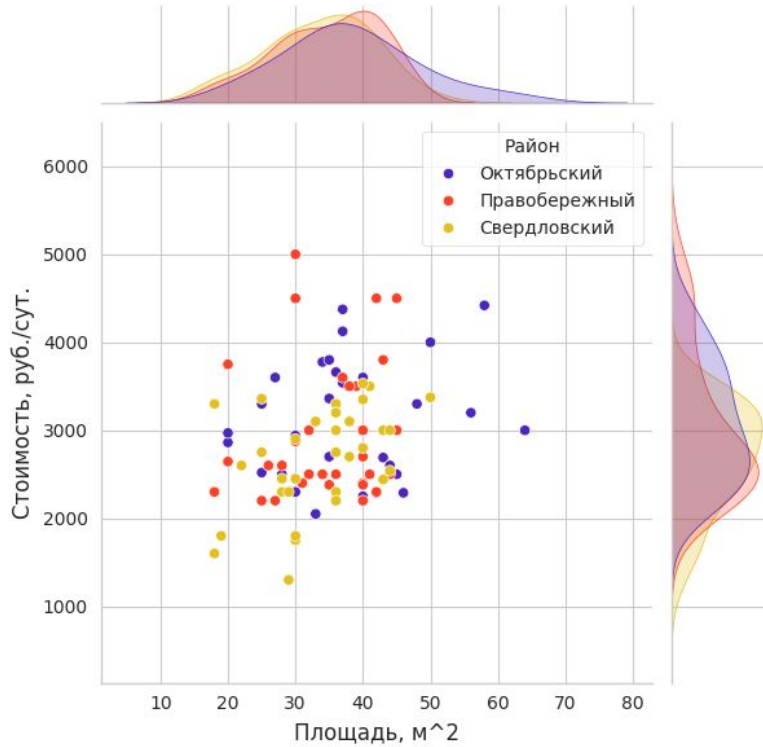
```
[63] df.groupby('Район')['Стоимость, руб./сут.'].describe()
```



	count	mean	std	min	25%	50%	75%	max
Район								
Октябрьский	34.00	3088.85	672.83	2050.00	2505.00	2985.00	3600.00	4420.00
Правобережный	33.00	3025.15	814.56	2200.00	2400.00	2645.00	3500.00	5000.00
Свердловский	33.00	2698.33	590.50	1300.00	2300.00	2750.00	3200.00	3530.00

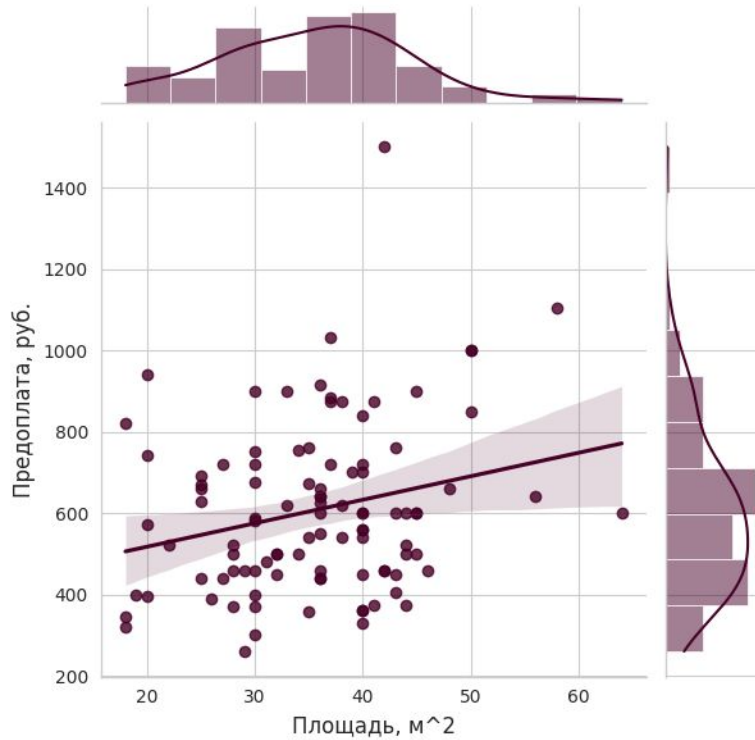


Визуализация данных

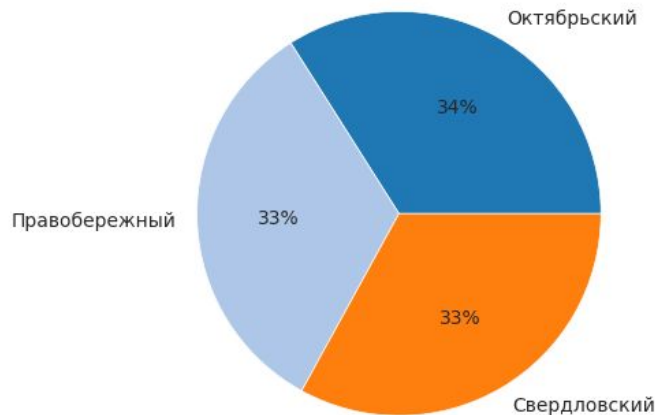


```
✓ [77] sns.jointplot(x = "Площадь, м^2", y = "Стоимость, руб./сут.", data = df, palette='CMRmap', hue='Район')  
сек.
```

Визуализация данных



```
sns.jointplot(x = "Предоплата, руб.", y = "Стоимость, руб./сут.", data = df, kind = 'reg', palette='CMRmap')
```



```
[133] sns.set_style('whitegrid')
      colors = sns.color_palette('tab20')[ 0:3 ]
      plt.pie(df['Район'].value_counts(), labels=['Октябрьский', 'Правобережный', 'Свердловский'], colors = colors, autopct='%0f%%')
      plt.show()
```

Python

Визуализация данных

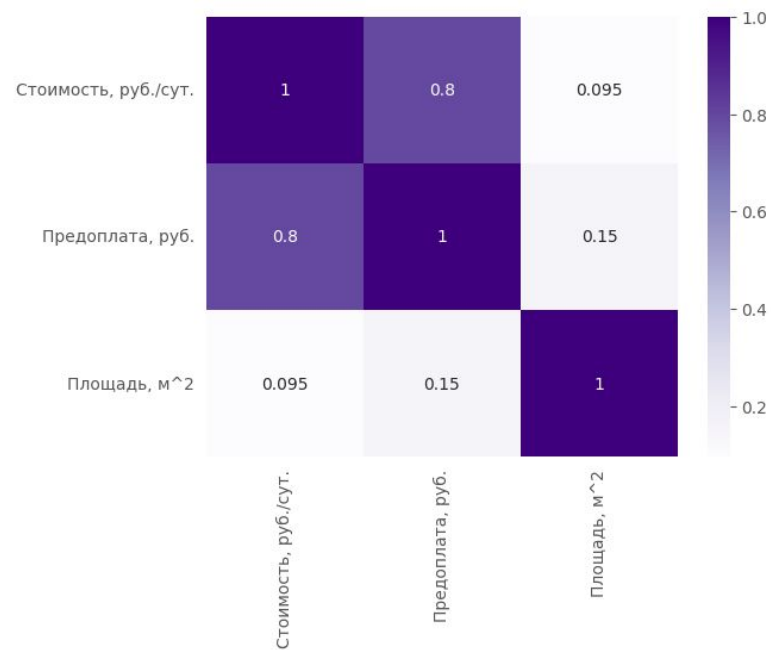
```
[108]
df_new = df[df['Район'] == 'Правобережный']
data = df_new[['Стоимость, руб./сут.', 'Предоплата, руб.', 'Площадь, м^2']]

print(data.corr())

dataplot = sns.heatmap(data.corr(), cmap="CMRmap", annot=True)
```

	Стоимость, руб./сут.	Предоплата, руб.	Площадь, м^2
Стоимость, руб./сут.	1.00	0.80	0.10
Предоплата, руб.	0.80	1.00	0.15
Площадь, м^2	0.10	0.15	1.00

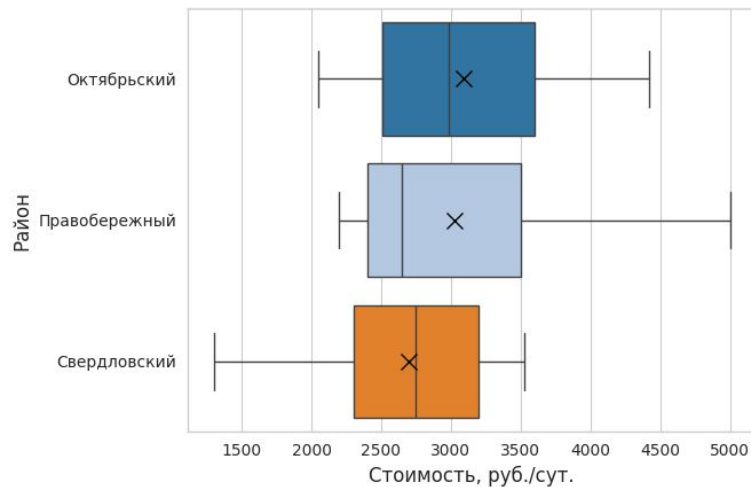
Python



Визуализация данных

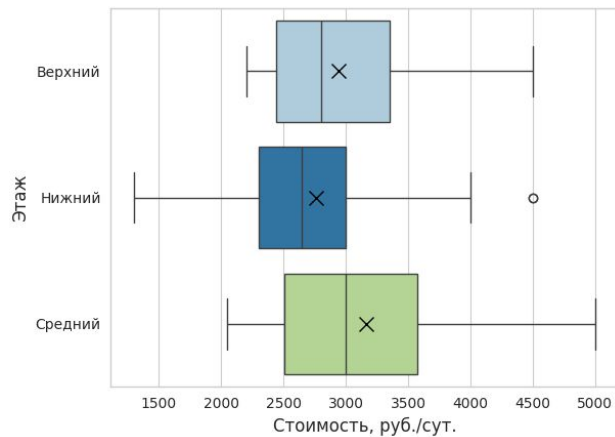
Python

Визуализация данных

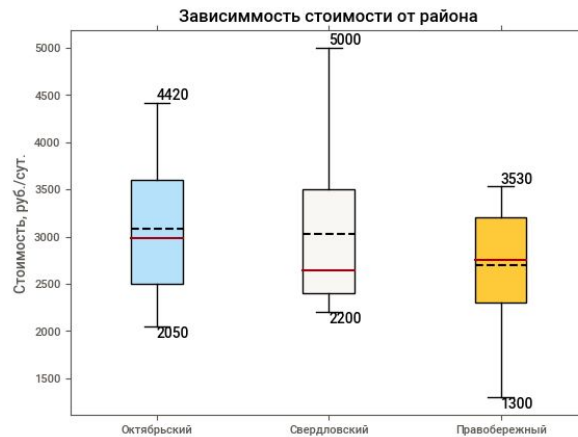


```
sns.boxplot(y='Район', x='Стоимость, руб./сут.', data=df[['Стоимость, руб./сут.', 'Район']],  
            palette = "tab20", showmeans=True,  
            meanprops={"marker": "x",  
                       "markeredgcolor": "black",  
                       "markersize": "10"}, orient = 'h')
```

Python



Визуализация данных



Python

Sweetviz

Библиотека на основе Pandas с открытым исходным кодом для выполнения основных задач EDA

Создает автономный HTML-отчет, который можно просматривать напрямую в браузере

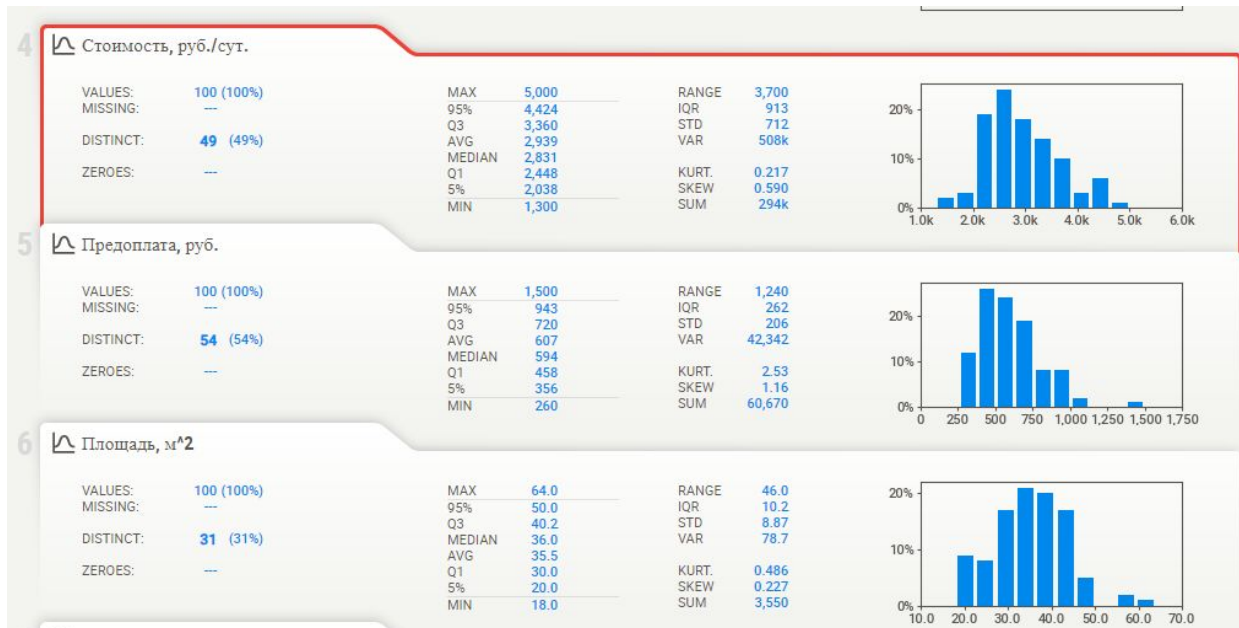
Возможности

Целевой анализ

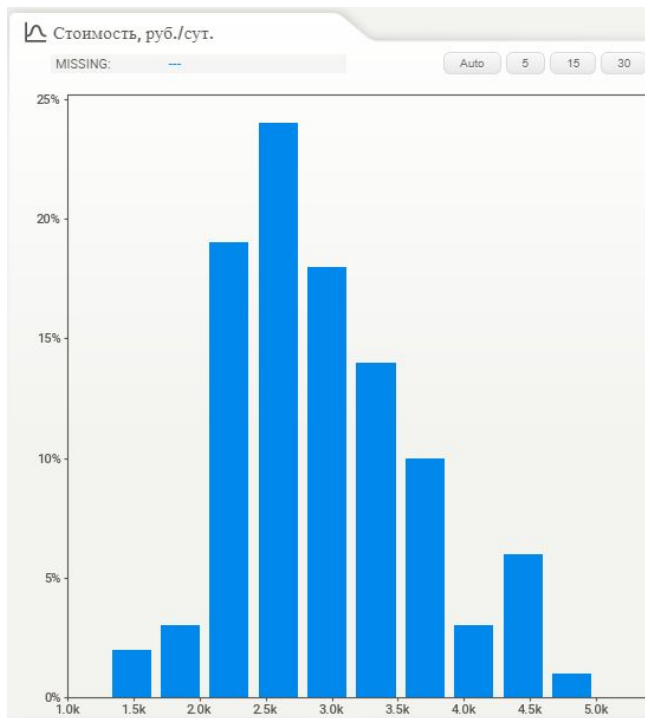
Сравнение наборов данных и групп внутри него

Выявление ассоциаций и корреляций

Подробная и быстрая визуализация



Python



Sweetviz



KNIME

www.knime.org

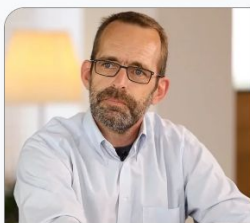
KNIME - это платформа с открытым исходным кодом, предназначенная для создания и выполнения различных рабочих процессов обработки данных



В 2004 году в университете Констанца на юге Германии команда разработчиков, специализирующейся на фармацевтических приложениях, начала работать над новой платформой с открытым исходным кодом в качестве инструмента для совместной работы и исследований

Первая версия KNIME была выпущена в июле 2006 года и была принята несколькими фармацевтическими компаниями

Сегодня пользователей KNIME можно найти во многих подразделениях крупных предприятий в широком спектре отраслей более чем в 60 странах



Michael Berthold

CEO



Thomas Gabriel

CHIEF OF STAFF



Peter Ohl

COMPLIANCE OFFICER



Bernd Wiswedel

CTO

<https://www.knime.com/team>

KNIME

Преимущества и возможности

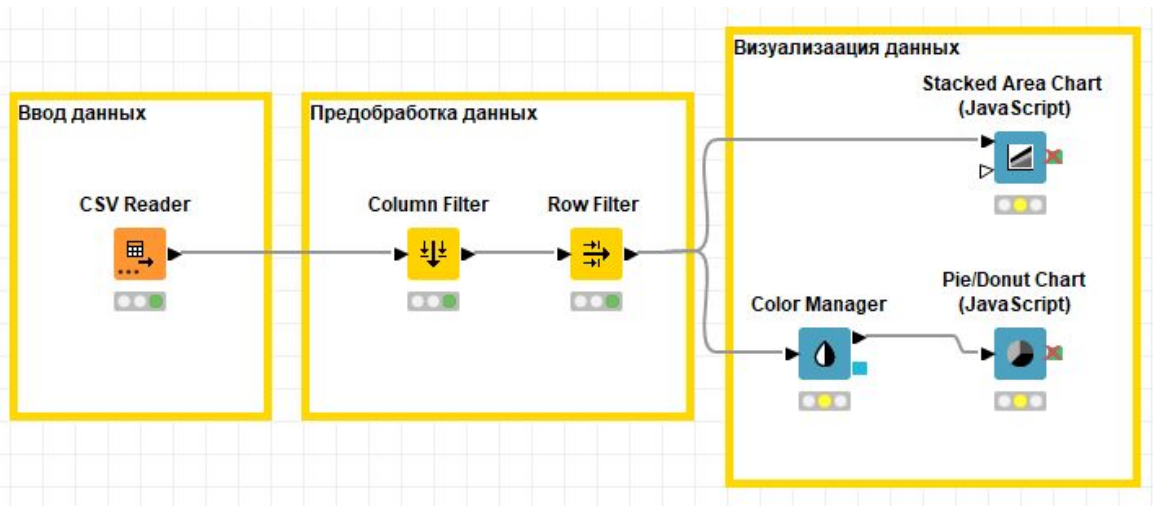
1. Интеграция данных из различных источников
2. Обработка и преобразование данных
3. Моделирование и анализ данных
4. Автоматизация процессов

И всё это без необходимости
написания кода!

KNIME

Принцип работы

Использует графический интерфейс для создания рабочих процессов (workflow), составленных из узлов (nodes), каждый из которых выполняет определенную операцию



KNIME

Описательные статистики

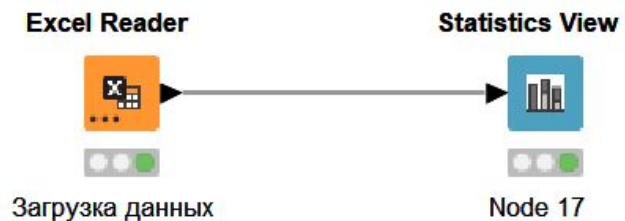
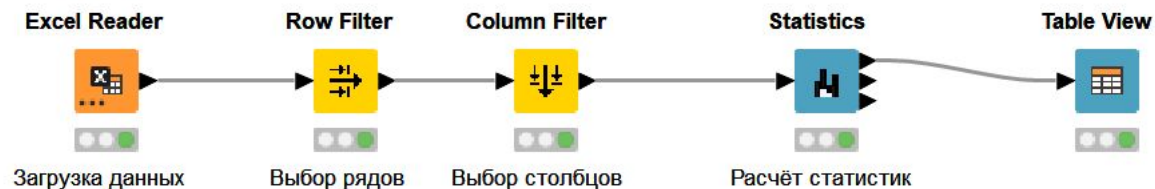
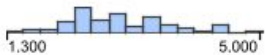
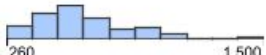



Table View

RowID	Min	Max	Mean	Std. deviation	Variance	Skewness	Kurtosis	Histogram
Стоимость, руб./сут.	1,300	5,000	2,943.192	714.8	510,938.993	0.574	0.192	
Предоплата, руб.	260	1,500	606.465	206.805	42,768.231	1.155	2.478	
Площадь, м^2	18	64	35.606	8.852	78.364	0.212	0.522	

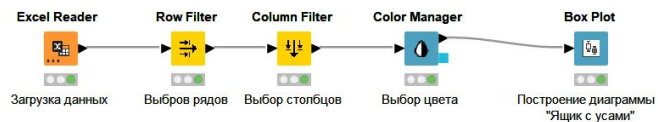
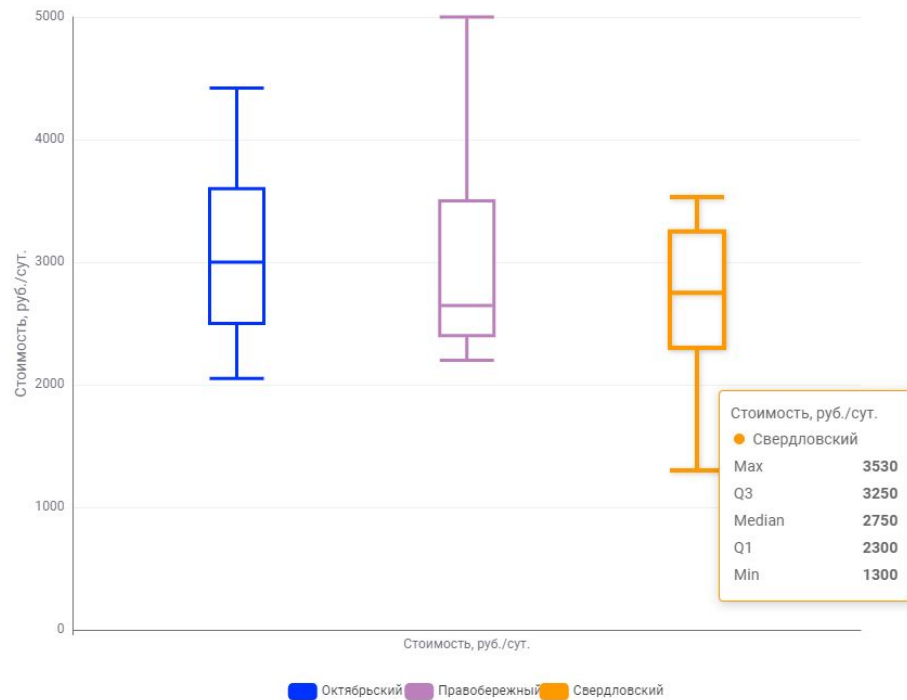
Statistics

Name	Minimum	Maximum	25% Quantile	50% Quantile (Median)	75% Quantile	Mean	Standard Deviation	Sum
Стоимость, руб./сут.	1,300	5,000	2,442.5	2,831	3,361.5	2,938.96	712.439	293,896
Предоплата, руб.	260	1,500	458	594	720	606.7	205.771	60,670
Площадь, м^2	18	64	30	36	40.75	35.5	8.871	3,550

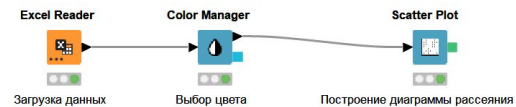
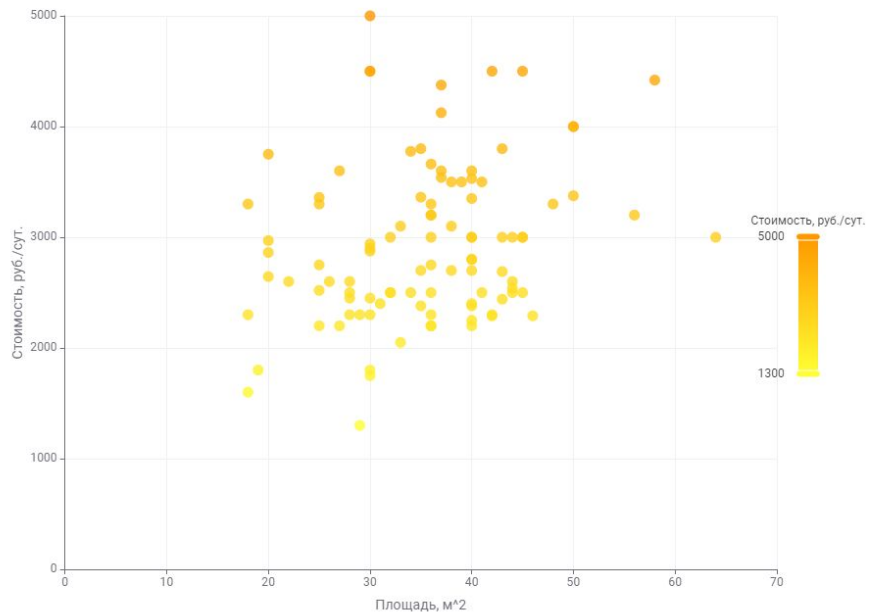
KNIME

Визуализация

Зависимость стоимости от района



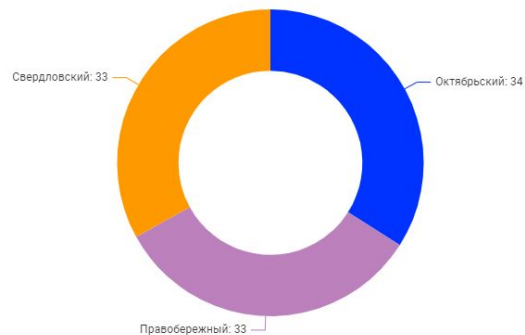
Зависимость стоимости от площади



KNIME

Визуализация

Pie Chart



⌵

Data

Category dimension

Район

Frequency dimension

Стоимость, руб./сут.

Aggregation

☐ None ☒ Occurrence count ☐ Sum

☐ Average

☒ Aggregate small categories

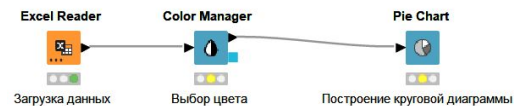
Threshold

3

Plot

Title

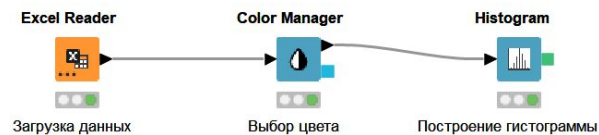
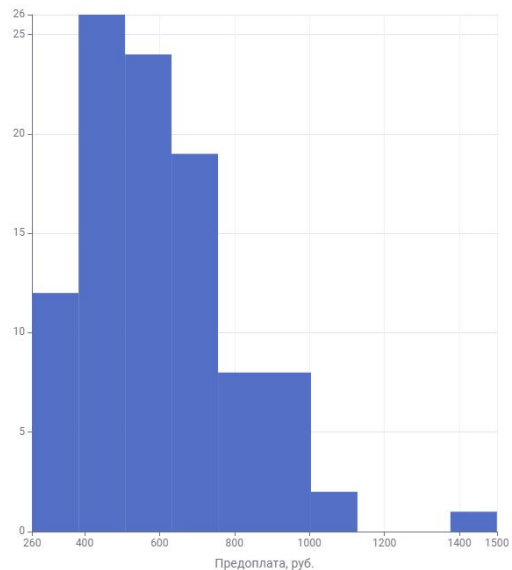
Pie Chart



KNIME

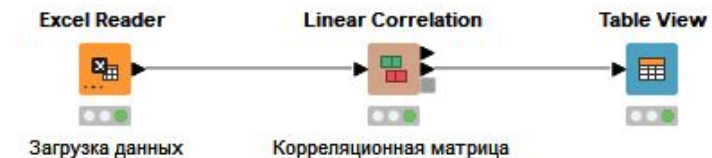
Визуализация

Histogram



KNIME

Корреляционный анализ



Стоимость, руб./сут.	Предоплата, руб.	Площадь, м^2
1	0.827	0.263
0.827	1	0.249
0.263	0.249	1

Row ID	Стоимость, руб./сут.	Предоплата, руб.	Площадь, м^2
Стоимость, руб.... 1		0.827	0.263
Предоплата, руб. 0.827	1		0.249
Площадь, м^2 0.263	0.249	0.249	1

<div><div></div> corr = -1</div> <div><div></div> corr = +1</div> <div><div></div> corr = n/a</div>	Стоимость, руб./сут.	Предоплата, руб.	Площадь, м^2
Стоимость, руб....			
Предоплата, руб.			
Площадь, м^2			

Выводы

Python

Большая гибкость

Большое сообщество
пользователей и
разработчиков

Требует базовых навыков
программирования

KNIME

Интуитивно понятный интерфейс

Подходит для специалистов без
навыков программирования

Быстрое получение результатов

Выбор между Python и KNIME будет зависеть от ваших конкретных задач и целей в области разведочного анализа данных

Источники

Python

https://dzen.ru/a/YFM4_gLK4zTIRduL

<https://habr.com/ru/articles/564172/>

KNIME

<https://dzen.ru/a/XhCU0HgSXgCx6jmp>

https://dzen.ru/a/XftzKy_ahvuENHLc

<https://questu.ru/articles/272435/>

<https://blog.knoldus.com/linear-regression-with-knime/https://blog.knoldus.com/linear-regression-with-knime/>

