МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное бюджетное образовательное учреждение высшего образования
«МИРЭА - Российский технологический университет»
Институт информационных технологий (ИИТ)
Кафедра прикладной математики
Направление «Прикладная информатика»
профиль «Управление данными»

Курсовая работа по дисциплине «Языки программирования для статистической обработки данных» по теме:

«Использование методов иерархической кластеризации и алгоритма -k means для определения факторов влияющих на объем выработки электроэнергии газовой турбиной»

Автор: студент группы ИНБО-22-23
Снигаренко Антон Владимирович
Руководитель: старший преподаватель
Трушин Степан Михайлович

Москва 2025

Объект исследования

В рамках исследования рассматривается применение передовых методов анализа данных, в частности иерархической кластеризации и алгоритма k- means, для идентификации и оценки значимости факторов, оказывающих влияние на объем генерации электрической энергии газовыми турбинами. Эффективность эксплуатации газотурбинных установок (ГТУ) напрямую зависит от множества операционных и внешних параметров. Целью работы является выявление скрытых закономерностей в многомерных массивах данных, регистрируемых в процессе работы ГТУ, с последующим формированием кластеров, характеризующих различные режимы работы и соответствующие им уровни выработки.

Описание используемых данных

В данной работе был использован синтетический датасет, который создан с помощью Python кода. Данный набор данных максимально приближен к реальным замерам. Набор можно увидеть в приложении А.

В наборе данных имеется 10 столбцов, 9 из которых численные и 1 категориальная.

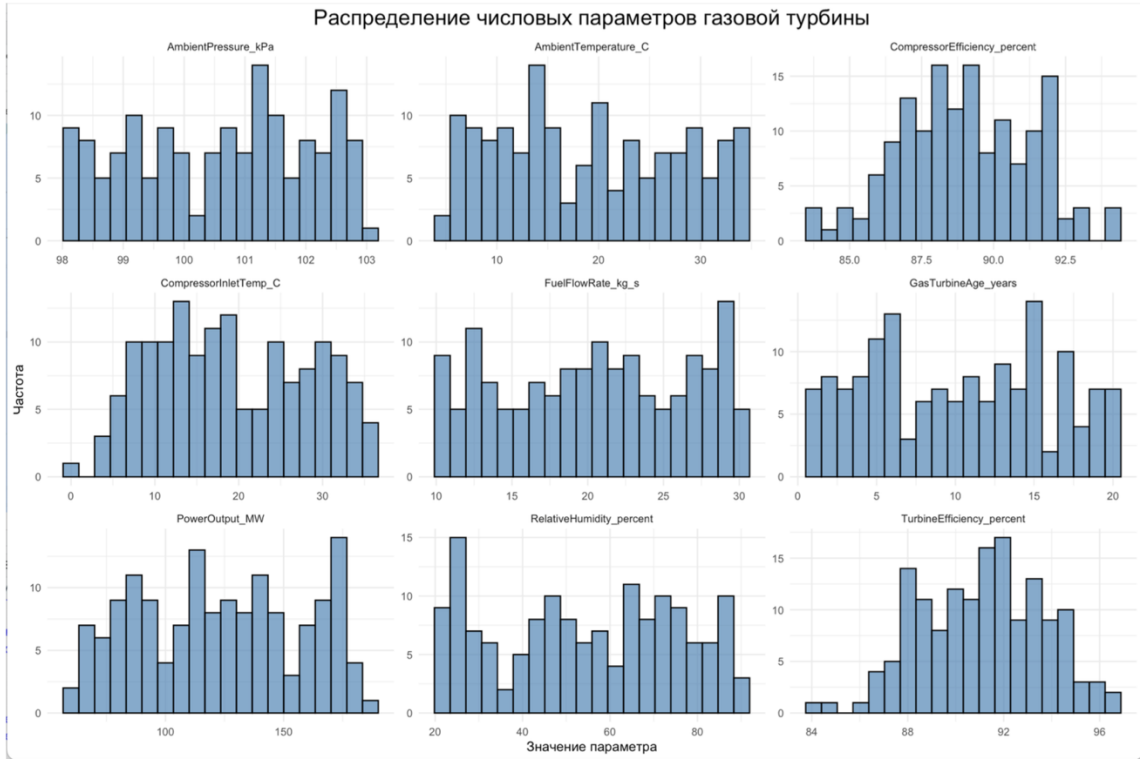
Все значения в датасете в большей или меньшей степени влияют на выработку электроэнергии.

Какие именно значения будут больше влиять – должны узнать с помощью подробного анализа данных

Основные методы анализа (EDA)

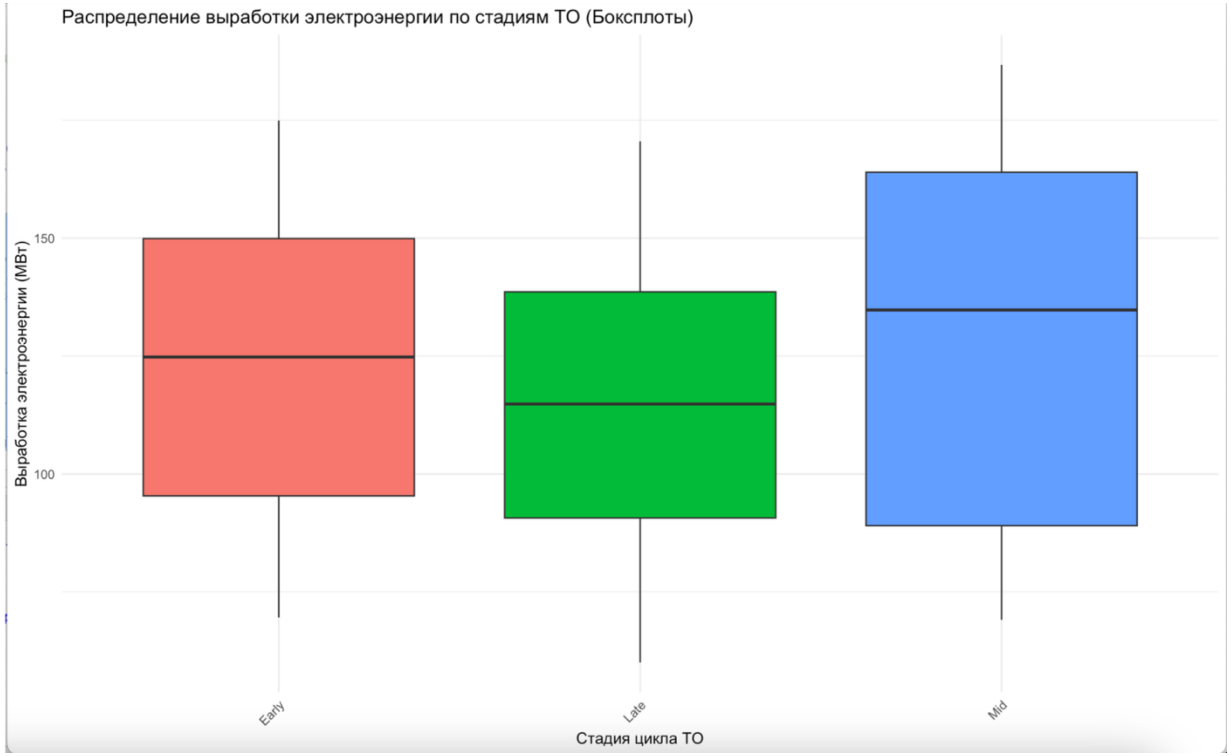
К основным методам анализа в данной работе можно отнести: Первичный анализ данных, корреляционный анализ, ANOVA и иерархическая кластеризация с помощью алгоритма k-means.

Результаты анализа



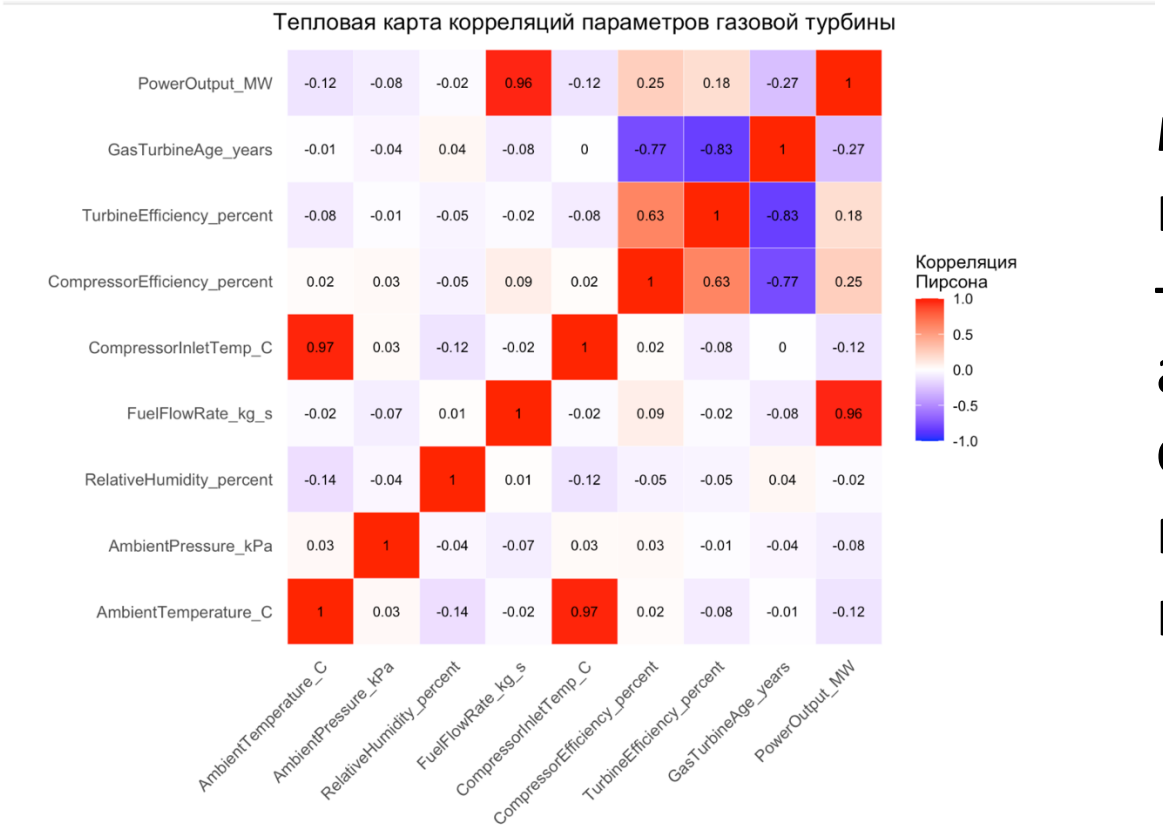
Приведем результат первичного анализа данных. На данных графиках мы можем проследить картину в общем и понять, какие методы нам подойдут лучше. В данном случае можно заметить мультимодальность и пиковые бугры в некоторых графиках

Боксплот по категориальным данным



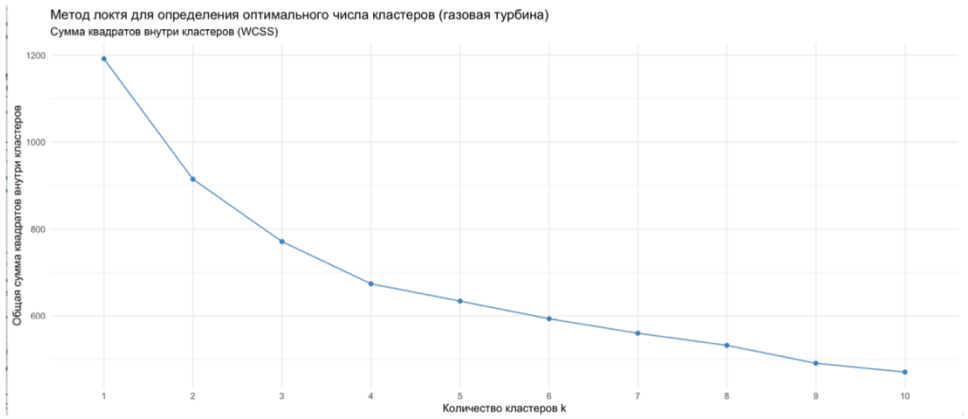
На данных графиках мы можем видеть разброс значений, в зависимости от времени после ТО.

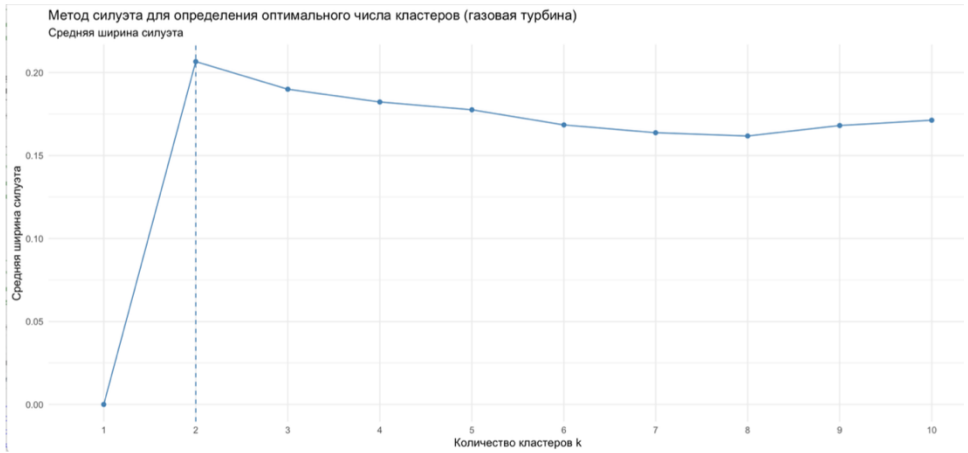
Корреляционный анализ данных



На данной тепловой карте мы можем увидеть корреляционный анализ данный – один из важнейших методов анализа в данной работе так как он показывает зависимости переменных (синие или красные не считая побочной диагонали)

Методы для определения количества кластеров





Метод силуэтов и метод локтя поможет нам определить количество кластеров и правильно провести анализ в будущем.

В рамках этой работы будет выбрано 3 кластера – как наглядная демонстрация работы алгоритмов

Кластеризация данных

Последним методом, использованном в рамках данной курсовой работы будет кластеризация данных, визуализация которой будет показана в Приложение Б

Выводы

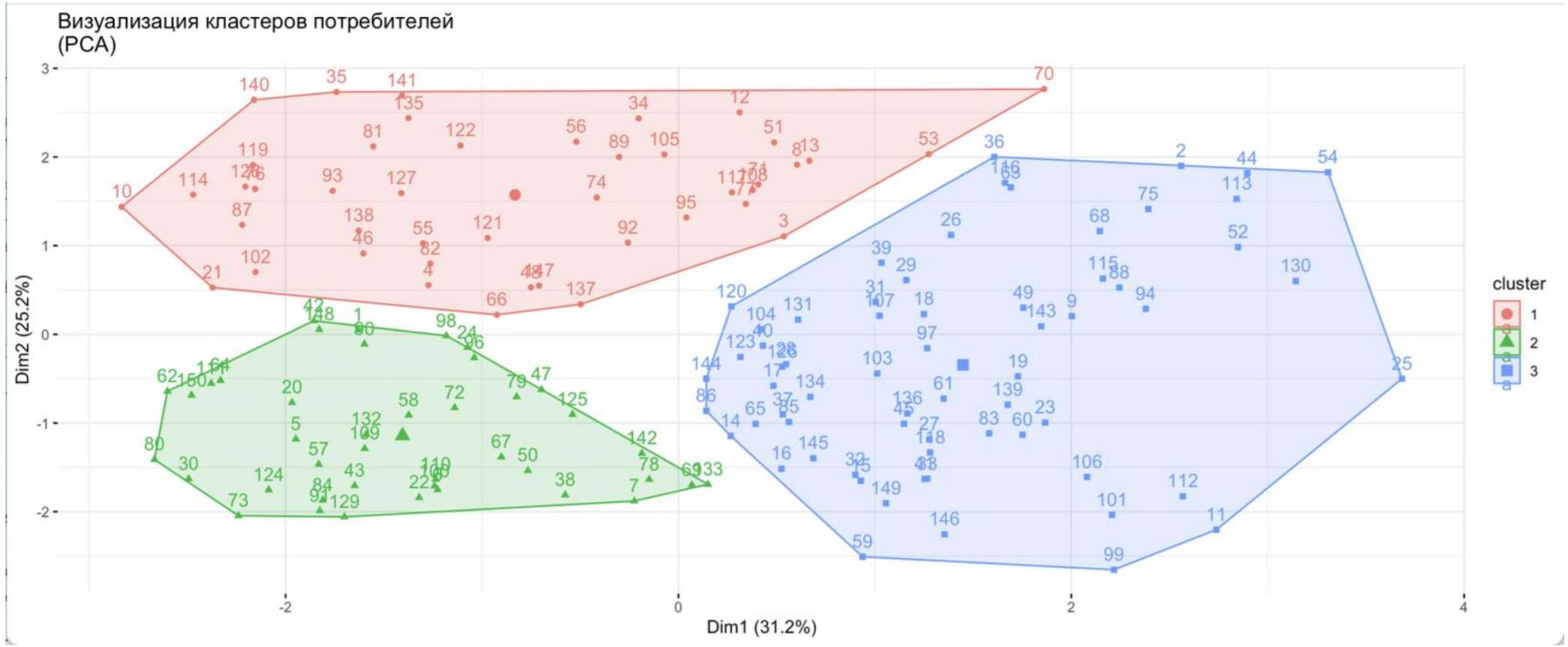
В ходе проведенного исследования были использованы методы иерархической кластеризации и алгоритм k-means для анализа факторов, влияющих на объем выработки электроэнергии газовой турбиной. Оба подхода показали свою применимость для решения поставленной задачи, позволив выделить группы режимов работы турбины с похожими характеристиками и определить ключевые параметры, оказывающие влияние на энергетическую эффективность.

Приложение А

AmbientTemperature_C	AmbientPressure_kPa	RelativeHumidity_percent	FuelFlowRate_kg_s	CompressorInletTemp_C	CompressorEfficiency_percent	TurbineEfficiency_percent	GasTurbineAge_years	MaintenanceCycle	PowerOutput_MW
16.2	102.54	23.6	25.54	17.0	91.89	92.04	5	Mid	158.06
33.5	99.2	57.2	21.17	32.3	86.32	87.0	18	Mid	116.48
27.0	98.72	57.8	18.48	27.1	88.22	90.07	10	Early	115.34
23.0	100.45	64.6	28.13	22.6	91.77	91.77	6	Early	174.07
9.7	102.93	70.8	12.22	12.0	90.64	93.64	1	Mid	86.16
9.7	99.21	88.3	19.85	10.2	88.39	94.78	5	Mid	125.51
6.7	101.36	56.1	10.23	7.4	86.97	93.53	9	Early	84.41
31.0	101.81	42.6	19.37	30.2	89.31	89.34	12	Late	105.8
23.0	99.19	75.7	11.13	22.1	85.88	88.15	14	Early	69.55
26.2	101.64	39.0	12.38	25.4	94.03	94.91	2	Early	95.51
5.6	99.84	50.7	12.35	6.4	83.78	87.37	17	Late	69.5
34.1	101.16	25.5	22.98	33.3	90.34	90.35	14	Late	130.63
30.0	101.17	21.8	24.92	30.6	88.42	90.1	13	Late	138.95
11.4	100.68	87.4	21.67	15.5	89.15	89.13	9	Late	116.71
10.5	98.45	78.5	29.24	12.2	86.68	91.32	15	Mid	174.65
10.5	102.18	68.7	17.5	9.9	88.1	91.69	15	Late	98.27
14.1	99.6	48.6	15.71	16.5	88.64	89.68	11	Early	102.34
20.7	98.93	32.1	27.37	19.9	88.85	88.02	16	Late	147.77
18.0	98.2	31.0	14.47	13.9	87.07	88.63	17	Late	79.48
13.7	100.95	37.5	29.26	11.7	91.03	93.9	4	Mid	181.45
23.4	101.39	58.4	10.24	19.6	91.15	95.59	1	Early	80.7
9.2	98.08	70.0	29.4	8.5	89.51	94.82	8	Mid	179.68
13.8	100.56	66.2	10.86	13.8	87.36	87.93	17	Late	63.64
16.0	99.13	39.6	27.82	19.3	88.53	92.6	3	Mid	167.4
18.7	101.23	86.8	20.55	19.3	83.98	84.93	20	Early	116.0
28.6	98.87	71.7	29.86	28.1	88.61	87.9	15	Early	174.93
11.0	101.45	58.8	11.48	12.6	87.76	88.95	15	Mid	76.43
20.4	99.93	62.8	21.08	16.0	87.52	92.09	14	Mid	123.08
22.8	102.68	49.4	29.39	23.2	87.83	88.62	14	Late	162.9
6.4	98.69	37.3	20.46	7.9	91.92	94.54	2	Early	140.71
23.2	99.71	44.9	22.59	20.3	87.29	89.78	13	Late	123.27

Датасет

Приложение Б



Итоговая кластеризация

Благодарю за внимание!