**Федеральное агентство по образованию Российской Федерации**

**Государственное образовательное учреждение**

**высшего профессионального образования**

**«КУБАНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»**

**Кафедра прикладной математики**

Преподаватель,

д.т.н. А.А. Халафян

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**ОТЧЕТ**

**СТАТИСТИЧЕСКИЙ АНАЛИЗ ЛУЧШИХ ДЕЙСТВУЮЩИХ ИГРОКОВ НБА**

Работу выполнил студент 4 курса

факультета компьютерных технологий и прикладной математики  
спец. 01.03.02 – Прикладная математика и информатика

Сторчак Вадим, группа 4ИТ

Краснодар 2021

**СОДЕРЖАНИЕ**

[**1 Исходные данные 4**](#_Toc87056809)

[**2 Графический анализ 6**](#_Toc87056810)

[**2.1 2D Graphs 6**](#_Toc87056811)

[**2.1.1 2D Histogramms 6**](#_Toc87056812)

[**2.1.2 2D Scatterplots 8**](#_Toc87056813)

[**2.1.3 2D Box Plots 10**](#_Toc87056814)

[**2.2 Средство «закрашивание» 11**](#_Toc87056815)

[**3 Основные статистики 14**](#_Toc87056816)

[**3.1 Описательные статистики 14**](#_Toc87056817)

[**3.2 Корреляционная матрица 14**](#_Toc87056818)

[**3.3 Критерий Стьюдента сравнения средних 15**](#_Toc87056819)

[**3.3.1 t-test, independent, by groups 15**](#_Toc87056820)

[**3.3.2 t-test, independent, by variables 16**](#_Toc87056821)

[**3.3.3 t-test, dependent samples 16**](#_Toc87056822)

[**3.3.4 t-test, single samples 16**](#_Toc87056823)

**3.**[**4 Группировка и однофакторная ANOVA 17**](#_Toc87056824)

[**4 Дисперсионный анализ 22**](#_Toc87056825)

[**5 Непараметрическая статистика 25**](#_Toc87056826)

[**6 Частотный анализ 27**](#_Toc233138317)

[**7 Линейное многомерное моделирование взаимосвязей 29**](#_Toc233138318)

[**8 Нелинейное многомерное моделирование взаимосвязей 32**](#_Toc233138319)

[**8.1 Fixed Nonlinear Regression 32**](#_Toc233138320)

[**8.2 Логит регрессия 33**](#_Toc233138321)

[**8.3 Пробит регрессия 35**](#_Toc233138322)

[**8.4 Экспоненциальная регрессия 36**](#_Toc233138323)

[**8.5 Кусочно-линейная регрессия 39**](#_Toc233138324)

[**8.6 Определенная пользователем регрессия 41**](#_Toc233138325)

[**9 Канонический анализ 43**](#_Toc233138326)

[**10 Дискриминантный анализ 48**](#_Toc233138327)

[**11 Классификационный анализ без обучения 52**](#_Toc233138328)

[**11.1 Кластерный анализ 52**](#_Toc233138329)

[**11.1.1 Метод к-средних 52**](#_Toc233138330)

[**11.1.2 Двухвходовая кластеризация 54**](#_Toc233138331)

[**11.1.3 Древовидная кластеризация 56**](#_Toc233138332)

[**11.2 Деревья классификации 58**](#_Toc233138333)

[**12 Методы редукции данных 63**](#_Toc233138334)

[**12.1 Факторный анализ 63**](#_Toc233138335)

[**12.2 Метод анализ главных компонент и классификация 67**](#_Toc233138336)

[**13 Многомерное шкалирование 74**](#_Toc233138337)

**1 Исходные данные**

Изображение выглядит как стол

Автоматически созданное описание

В исходной таблице представлены данные по действующим лучшим игрокам баскетбольной лиги NBA.

Показатели:

1. Player – имя и фамилия игрока.
2. Age – возраст игрока на текущий момент.
3. Цвет кожи – цвет кожи игрока;
4. Национальность – национальность игрока, за какую сборную играет;
5. TM – команда, за которую играет игрок.
6. Pos – позиция, под которой играет игрок.
7. G – количество игр, сыгранных игроком на момент сбора статистики.
8. GS – количество игр, в которых игрок выходит в стартовой пятерке.
9. MP – количество сыгранных минут.
10. FG – количество забитый бросков.
11. FGA – количество совершённых попыток.
12. FG% – общий процент попаданий игровых бросков.
13. 3P – количество реализованных трёх очковых бросков;
14. 3PA – количество совершённых трёх очковых попыток;
15. 3P% – процент реализованных трёх очковых бросков;
16. 2P – количество забитых двух очковых бросков;
17. 2PA – количество совершённых двух очковых попыток;
18. 2P% - процент реализованных двух очковых бросков;
19. FT – количество забитых штрафных бросков.
20. FTA – количество совершённых попыток штрафного броска.
21. FT% - процент попадания со штрафной линии.
22. ORB – количество подборов совершённых на щите соперника.
23. DRB – количество подборов совершённых на своё щите.
24. TRB – общее количество подборов.
25. AST – количество отданных результативных передач.
26. STL – количество отборов мяча у соперника.
27. BLK – количество заблокированных бросков.
28. TOV – количество потерь игрока.
29. PF – количество персональных замечаний игрока.
30. PTS – количество набранных очков.
31. Trp Dbl – количество триплдаблов (когда игрок набирает 10 и более в трёх статистически полезных показателях: PT, AST, REB, BLK, STL. В рамках одного матча).

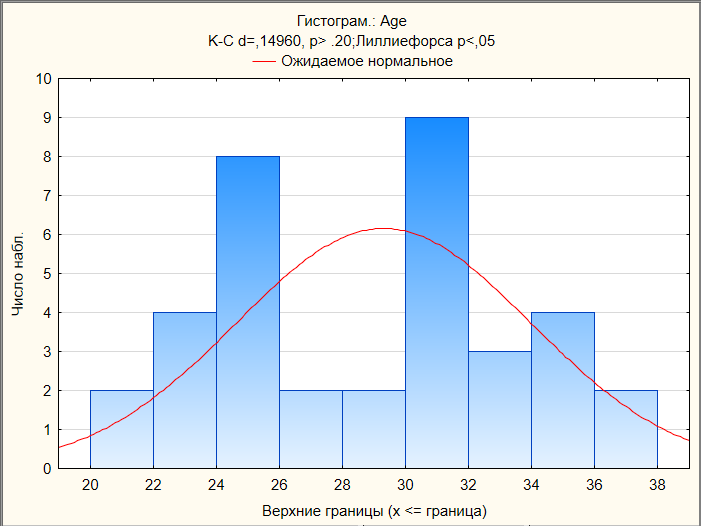
**2 Графический анализ**

**2.1 2D Graphs**

**2.1.1 2D Histogramms**

2D Histogramms являются графическими представлениями распределения частот выбранных переменных.

2D Histogramms Regular (простые) – столбчатая диаграмма распределения частот для выбранной переменной.



Гистограмма была построена по количественной переменной Age (возраст) и показывает, что в генеральной совокупности чаще встречаются игроки возраста 30-32 лет (9 игроков) и возраста 24-26 (8 игроков), что составляет 25% и 22,22% соответственно.

2D Histogramms Multiple (составные) – изображают распределение частот для нескольких переменных на одном графике.



Данная гистограмма построена по общему числу результативных передач и общему числу подборов у игроков. Из гистограммы видно, что большинство игроков имеют от 1000 до 2000 общих подборов и передач. Обусловлено это двумя причинами:

* формирование выборки;
* кол-во игроков, которые отыграли большое число сезонов в лиге крайне мало, так как такие игроки являются звездами и их по определению не может быть много.

2D Histogramms Double-Y (с двойной осью Y) – Комбинация двух по-разному масштабированных составных гистограмм.



Данная гистограмма построена по двум разномасштабным переменным: общее количество бросков и кол-во набранных очков. Из диаграммы видно, что чаще всего встречаются игроки набравшие от 5к до 10к очков, а так же совершившие от 10к до 15к, и от 0 до 5к попыток броска.

**2.1.2 2D Scatterplots**

2D Scatterplots (диаграммы рассеяния) визуализируют зависимость между двумя переменными.

2D Scatterplots Regular (протые) - визуализируют зависимость между двумя переменными X и Y.



Данная диаграмма отображает зависимости между количеством выброшенных бросков и набранными очками за карьеру с линейной подгонкой.

2D Scatterplots Multiple (составные) – состоит из нескольких зависимостей и изображает несколько корреляций.



Данная диаграмма отображает зависимости между кол-вом сыгранных матчей и числом забитых очков, а так же между кол-вом сыгранных матчей и кол-вом совершенных попыток атаки кольца.

2D Scatterplots Double-Y (с двойной осью Y) – комбинация двух составных диаграмм рассеяния для одной переменной X и двух различных наборов переменных Y.



Данная диаграмма отображает комбинацию двух составных диаграмм: первая зависимость между количеством реализованных трёхочковых бросков и количеством сыгранных игр, вторая зависимость между количеством реализованных двухочковых бросков и количеством сыгранных игр.

**2.1.3 2D Box Plots**

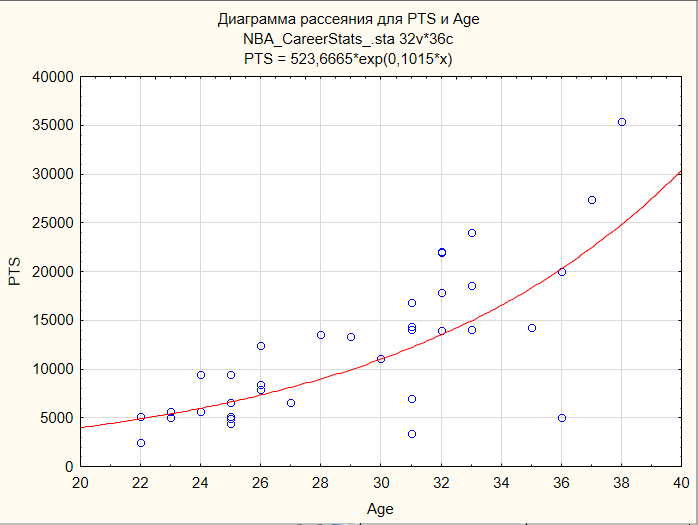
2D Box Plots (графики ящика – диаграммы размаха) – на этих диаграммах изображаются диапазоны или характеристики распределения значений выбранной переменной отдельно по группам, заданным категориальной переменной.



На диаграмме представлены характеристики распределения значений общих набранных очков по позициям игроков.

**2.2 Средство «закрашивание»**

Построим диаграмму рассеивания для переменных количества набранных очков за карьеру и возраста игрока.



Из диаграммы видно, что есть точки, которые располагаются далеко от линии регрессии, поэтому аппроксимация может быть некачественной. Для исключения этой ситуации воспользуемся средством «закрашивание», чтобы исключить некоторые точки.

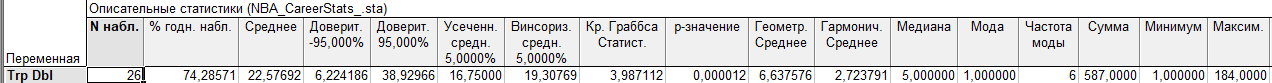


Сравнивая r до и после исключения точки, можно сказать, что качество аппроксимации увеличилось (доля объясненной дисперсии).

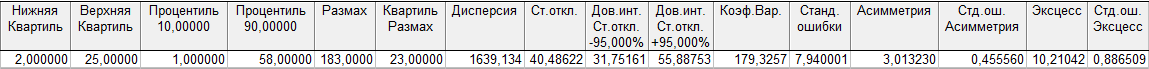
**3 Основные статистики**

**3.1 Описательные статистики**

Таблица описательных статистик для количества сделанных трипл даблов:



*продолжение*



В таблице представлены следующие статистики: число наблюдений 26, среднее 22.57692, сумма 587, медиана 5, геометрическое среднее 6.637576, гармоническое среднее 2.723791, стандартное отклонение 2.423481, дисперсия 1639.134, стандартная ошибка среднего 7.940001, доверительные пределы для среднего 6.224186 и 38.92966, ассиметрия 3.013230, стандартная ошибка ассиметрии 0.4555560, эксцесс 10.21042, стандартная ошибка эксцесса 0.886509, минимум 1, максимум 184, нижний квартиль 2, верхний квартиль 25, 10-я процентиль (квантиль 0,1) 1, 90- процентиль (квантиль 0,9) 58, размах 183, квартильный размах 23.

**3.2 Корреляционная матрица**

Изображение выглядит как текст, кроссворд

Автоматически созданное описание

Из таблицы видно, что сильная корреляция r>0.75 присутствует между средними количеством сыгранных игр и всем показателями кроме результативных передач, сыгранными минутами и всеми показателями, количеством реализованных бросков и всеми показателями кроме результативных передач, между общим числом подборов и сыгранных игр, сыгранных минут, совершенных бросков, а так же между возрастом игрока и общим числом подборов и сыгранных игр, сыгранных минут; средняя корреляция между количеством сыгранных матчей и результативными передачами, реализованных бросков и результативными передачами, результативными передачами и всеми другими показателями. Слабой корреляции не наблюдается. Красным отмечены значения, где уровень значимости <0.05.

## 3.3 Критерий Стьюдента сравнения средних

**3.3.1 t-test, independent, by groups**

Проверим равенство средних для количества сыгранных игр, числу сыгранных минут и количеству набранных типл даблов, по группам игроков на позиции разыгрывающего и малого форварда на основе t-критерия (t-test, independent, by groups).

Изображение выглядит как стол

Автоматически созданное описание

Как видно из таблицы все p>0.05, принимаем гипотезу о равенстве средних, то есть отличие средних статистически не значимо. Это связано с тем, что количество сыгранных минут и проведённых матчей зависит в первую очередь от таланта и здоровья игрока, но никак не от позиции, а вот для параметра количества трипл даблов недостаточная выборка (так как опираясь на собственный опыт, могу сказать, что количество триплдаблов должно быть сильно больше у разыгрывающих и атакующих защитников, если брать большую выборку). Для корректности вывода отличие дисперсий не должно быть статистически значимым (p дисп > 0.05). Соответственно, относительно всех выбранных переменных в рамках выбранных групп вывод корректен.

**3.3.2 t-test, independent, by variables**

Метод используется для сравнения двух параметров одной величины для различных объектов. Так как в исходной таблице не имеются подходящие данные, найдём данные по общему проценту попаданий команд, в которых играют игроки и добавим этот столбец TM\_FG% в таблицу. Проведём проверку гипотезы о равенстве средних для параметров FG% и TM\_FG%.

Изображение выглядит как стол

Автоматически созданное описание

Из таблицы видно, что гипотеза о равенстве средних отвергается, так как p>0.05. Следовательно можно сказать, что средние значения FG% игрока и команды, в которой он играет статистически не значимо.

**3.3.3 t-test, dependent samples**

Проверка производится для при двух повторных измерениях случайных величин двух зависимых групп. Так как в таблице не имеется подходящих данных следует добавить параметр, по любому из параметров, который был на момент прошлого сезона. Например, процент 3-х очковых бросков – Of\_Last\_Season\_3P%. Тогда можно проверить гипотезу равенства средних для добавленного параметра и 3P%.

Изображение выглядит как стол

Автоматически созданное описание

По данным таблицы уровень значимости меньше 0.05, поэтому гипотезу о равенстве средних отвергаем, отличие средних статистически значимо. Также, так как отличие статистически значимо, мы можем перенести вывод на генеральную совокупность, то есть в среднем процент 3-х очковых попаданий изменяется от сезона к сезону.

**3.3.4 t-test, single samples**

Предположим, что в среднем за карьеру игроки совершают 8000 попыток броска за карьеру. Проверим гипотезу, что среднее количество попыток броска за карьеру из выборки не отличается статистически значимо от указанного среднего.

Изображение выглядит как текст, стол

Автоматически созданное описание

По данным таблицы можно сделать вывод, что отличие статистически не значимо, так как p>0.05.

**3.4 Группировка и однофакторная ANOVA**

Для сравнения средних в более чем 2 группах необходимо воспользоваться модулем дисперсионного анализа ANOVA. Метод тоже параметрический и требуется выполнение тех же условий, что и для критерия Стьюдента (нормальный закон распределения переменной, гипотеза о равенстве дисперсий).

Для анализа выберем зависимые переменные: сыгранные минуты, трёхочковые броски, двух очковые броски, штрафные броски, а группирующая переменная – национальность.

Изображение выглядит как текст, кроссворд

Автоматически созданное описание

Итоговая таблица средних показывает среднее, количество наблюдений и стандартное отклонение сыгранных минут, трёхочковых бросков, двухочковых бросков и штрафных бросков для представителей каждой национальности. Необходимо проверить верна ли гипотеза о равенстве средних в генеральной совокупности.

Изображение выглядит как стол

Автоматически созданное описание

На вкладке дисперсионный анализ проверяется гипотеза о равенстве дисперсий для сыгранных минут, трёхочковых бросков, двухочковых бросков и штрафных бросков. Как видно из таблицы, уровень значимости p везде больше 0.05, поэтому верна гипотеза о равенстве дисперсий. Поэтому выполняется условие применимости однофакторного дисперсионного анализа.

Также можем рассмотреть критерий Левена и критерий Брауна и Форсайта.

Изображение выглядит как стол

Автоматически созданное описание

Изображение выглядит как стол

Автоматически созданное описание

Опять-таки p больше 0.05 говорим, что гипотеза о равенстве дисперсий верна.

Изображение выглядит как стол

Автоматически созданное описание

Изображение выглядит как стол

Автоматически созданное описание

По национальности USA посмотрим коэффициенты корреляции и уровни значимости. p < 0.05 для всех параметров.

Для проверки есть ли статистически значимые отличия средних генеральной совокупности в группирующих переменных возьмем критерий НЗР, Шеффе и Тьюки.

Изображение выглядит как стол

Автоматически созданное описание

Изображение выглядит как текст, кроссворд

Автоматически созданное описание

Изображение выглядит как стол

Автоматически созданное описание

Все критерии Шеффе, Тьюки и НЗР на всех сравнениях выдали значения p > 0.05, поэтому можем сказать, что верна гипотеза о равенстве средних всех генеральных совокупностей. То есть национальность не влияет на основные статистические показатели игрока

Выберем теперь 2 группирующие переменные – национальность и цвет кожи.

Изображение выглядит как стол

Автоматически созданное описание

Итоговая таблица средних содержит информация о среднем, количестве наблюдений и стандартном отклонении для сыгранных минут, трёхочковых бросков, двух очковых бросков и штрафных бросков в зависимости от национальности и цвета кожи.

Изображение выглядит как текст, стол

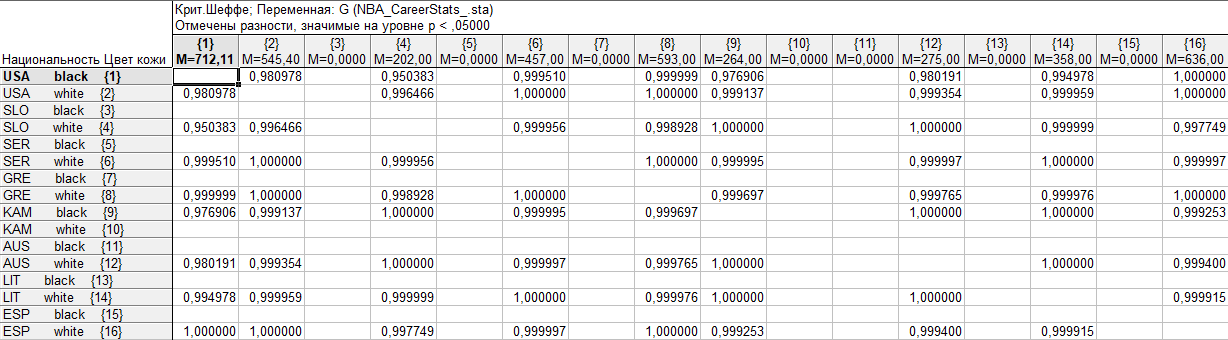
Автоматически созданное описание

Как видно из таблицы дисперсионного анализа, уровень значимости p везде больше 0.05, поэтому верна гипотеза о равенстве дисперсий. Поэтому выполняется условие применимости однофакторного дисперсионного анализа.

Рассмотрим теперь критерии НЗР, Шеффе и Тьюки.

Изображение выглядит как стол

Автоматически созданное описание



Изображение выглядит как текст, снимок экрана, полный, несколько

Автоматически созданное описание

Все критерии Шеффе, Тьюки и НЗР на всех сравнениях выдали значения p > 0.05, поэтому можем сказать, что верна гипотеза о равенстве средних всех генеральных совокупностей.

**4 Дисперсионный анализ**

Дисперсионный анализ является наиболее общим методом сравнения средних. В дисперсионном анализе можно исследовать зависимость количественного признака (зависимой переменной) от нескольких качественных признаков (факторов).

В качестве зависимых переменных выбраны количество подборов и результативных передач, а в качестве категориальных позиция игрока.

Изображение выглядит как стол

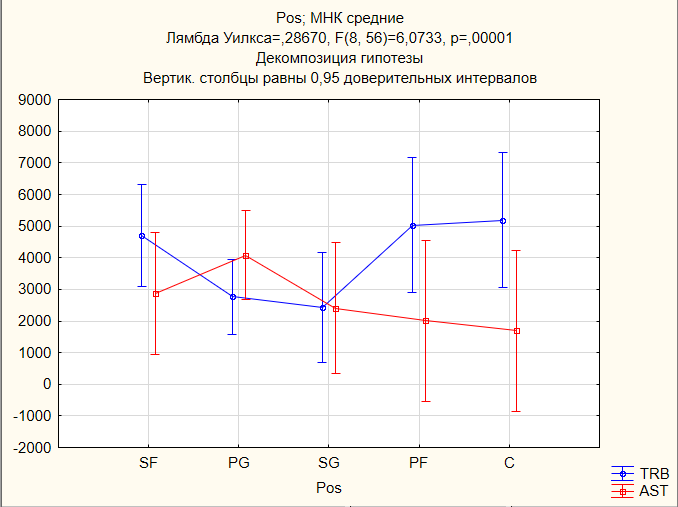
Автоматически созданное описание

В последнем столбце отображены уровни значимости критерия Фишера. Так как уровень значимости p < 0.05, то при разбиении на группы по позиции наверняка будут группы со статистически значимыми отличиями средних.

Изображение выглядит как текст, кроссворд

Автоматически созданное описание

На таблице МНК средние можем увидеть среднее, стандартную ошибку и доверительный интервал для подборов и результативных передач. Также эти данные можем посмотреть на графике.



Для применения дисперсионного анализа необходимо нормальное распределение зависимой переменной и равенство дисперсий. Посмотрим критерий Левена для игровых позиций.

Изображение выглядит как стол

Автоматически созданное описание

Как оказалось гипотеза о равенстве дисперсий верна для подборов и результативных передач. Посмотрим также диаграммы рассеяния.





Точки не находятся на одной прямой, а соответственно средние и стандартные отклонения коррелируют незначительно.

Изображение выглядит как стол

Автоматически созданное описание

Изображение выглядит как стол

Автоматически созданное описание

Изображение выглядит как стол

Автоматически созданное описание

Изображение выглядит как стол

Автоматически созданное описание

Изображение выглядит как стол

Автоматически созданное описание

Уровень значимости внутригрупповых корреляций, для каждой категории группирующей переменной.

# 5 Непараметрическая статистика

Если условия применения параметрических критериев (нормальность распределения, равенство дисперсий) не выполнены, необходимо воспользоваться непараметрическими критериями.

Коэффициент Спирмена используется, если переменные количественные (закон распределения переменных неизвестен или не является нормальным) и (или) качественные (порядковые). Рассмотрим переменные подборы в нападении, потери, блокшоты и персональные замечания.

Изображение выглядит как стол

Автоматически созданное описание

Корреляция для всех пар переменных умеренная (Спирмен R в промежутке 0.25–0.75), кроме подборов в нападении и персональных замечаний, там сильная >0.75. Так как p<0.05 во всех наблюдениях, то корреляция между переменными статистически значима.

Коэффициент Кендалла используется, если хотя бы одна переменная качественная (порядковая). Рассмотрим возраст игрока и количество набранных триплдаблов.

Изображение выглядит как стол

Автоматически созданное описание

У возраста игрока и числу набранных трипдаблов слабая корреляция (Кендалл тау <0.25). Так как p>0.05, то корреляция между переменными статистически не значима.

Коэффициент Гамма используют, если переменные содержат много повторяющихся значений.

Изображение выглядит как текст, стена, снимок экрана

Автоматически созданное описание

Для возраста игрока и числа набранных трипдаблов корреляция слабая (Гамма <0.25).

Ранговый ДА Фридмана и конкордация Кендалла используются для двух зависимых переменных (измерения которых производились более 2-х раз). Добавим параметр Start\_Last\_Season\_3P% и проверим критерий для нового параметра, Of\_Last\_Season\_3P% и 3P%.

Изображение выглядит как текст, внутренний, снимок экрана

Автоматически созданное описание

Коэффициент p < 0.05, значит верна альтернативная гипотеза о неравенстве средних в 3 группах.

Посмотрим попарно в каких группах средние неравны.

Изображение выглядит как стол

Автоматически созданное описание

При попарном сравнении всюду p < 0.05, значит верна гипотеза о неравенстве средних.

Теперь проверим, есть ли различие в цвете кожи, в зависимости от подборов, передач и набранных очков. Посмотрим сравнение нескольких независимых групп.

Изображение выглядит как текст, стол

Автоматически созданное описание

Изображение выглядит как текст, кроссворд

Автоматически созданное описание

Из таблиц мы делаем вывод, что верна гипотеза о равенстве средних в 2 группах (уровни значимости p больше 0.05).

Изображение выглядит как текст, внутренний, снимок экрана

Автоматически созданное описание

Изображение выглядит как текст, внутренний, снимок экрана

Автоматически созданное описание

Изображение выглядит как текст, внутренний, снимок экрана

Автоматически созданное описание

Уровни значимости для критерия сравнения средних для всех зависимых переменных.

Проверим, есть ли различие в возрасте в зависимости цвета кожи. Для этого проведем сравнение 2 независимых групп.

Изображение выглядит как стол

Автоматически созданное описание

Изображение выглядит как стол

Автоматически созданное описание

Изображение выглядит как стол

Автоматически созданное описание

Как результат, гипотеза о равенстве средних подтверждается, так как коэффициент p > 0.05.

**6 Частотный анализ**

**6.1 Таблицы частот**

Таблицы частот позволяют анализировать категориальные переменные.

Изображение выглядит как текст, кроссворд, шкафчик, табло

Автоматически созданное описание

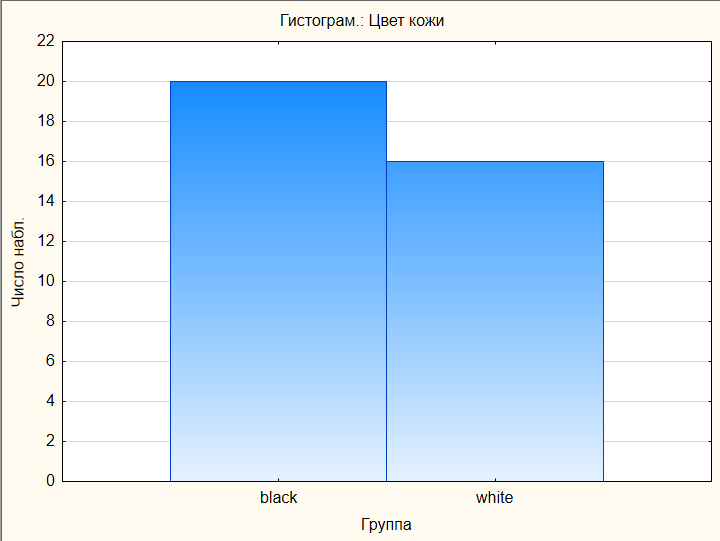
Данная таблица частот построена для категориальной переменной Pos. В таблицы для категорий SF, PG, SG, PF, C указаны слева направо: частоты, кумулятивные частоты, проценты (относительные частоты), кумулятивные проценты, 100%-кумулятивные проценты, логит преобразование частот, пробит преобразование частот.

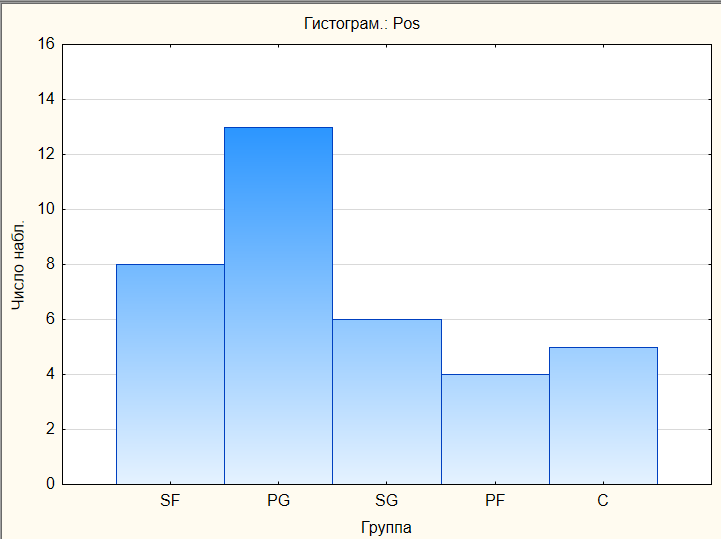
Изображение выглядит как стол

Автоматически созданное описание

Данная таблица частот построена для категориальной переменной Цвет\_кожи. В таблицы для категорий. black и white указаны те же показатели, что и в предыдущей таблице.

Гистограммы соответствующих категориальных переменных:





Из гистограмм видно, что больше игроков на игровой позиции разыгрывающего (PG), и игроков с черным цветом кожи.

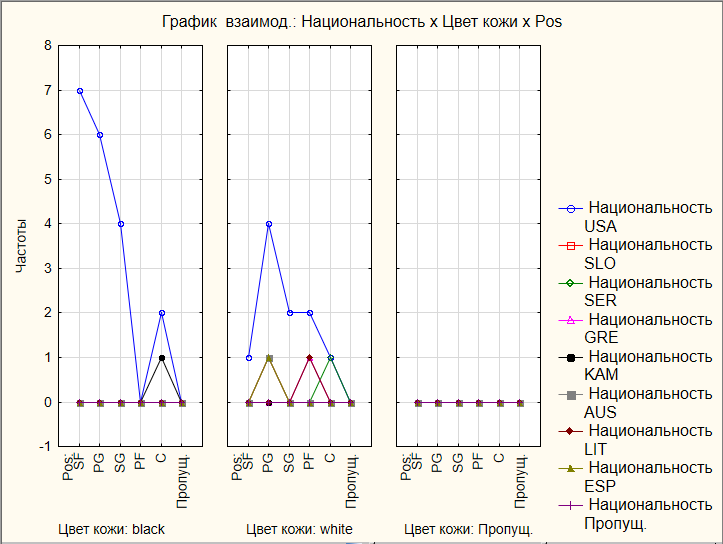
**6.2 Таблицы кросс табуляции.**

Таблицы кросстабуляции объединяют несколько таблиц частот. Например, по трём категориальным переменным – Национальность, Pos, Цвет кожи.

Изображение выглядит как стол

Автоматически созданное описание

Так же можно построить график взаимодействия частот.



**7 Линейное многомерное моделирование взаимосвязей**

Построим зависимость между Trpl\_Dbl по основным положительным характеристикам.

Изображение выглядит как стол

Автоматически созданное описание

Из таблицы видно, что коэффициента регрессии при переменных AST и STL статистически значимы (p-level<0.1), остальные статистически незначимы. Зависимость между откликом и предикторами не высокая (R2=0.5309).

Изображение выглядит как стол

Автоматически созданное описание

В целом регрессия значима.

Изображение выглядит как текст, кроссворд, внутренний, табло

Автоматически созданное описание

По данной таблице можно ранжировать влияния предикторов на отклик. Больше всего влияет AST. Этот же предиктор имеет наибольшую частную и получастную корреляции. R-square говорит о множественной корреляции между данной переменной и остальными переменными. Видно, что все коэффициенты детерминации умеренные.

Tolerance= 1 - R-square.

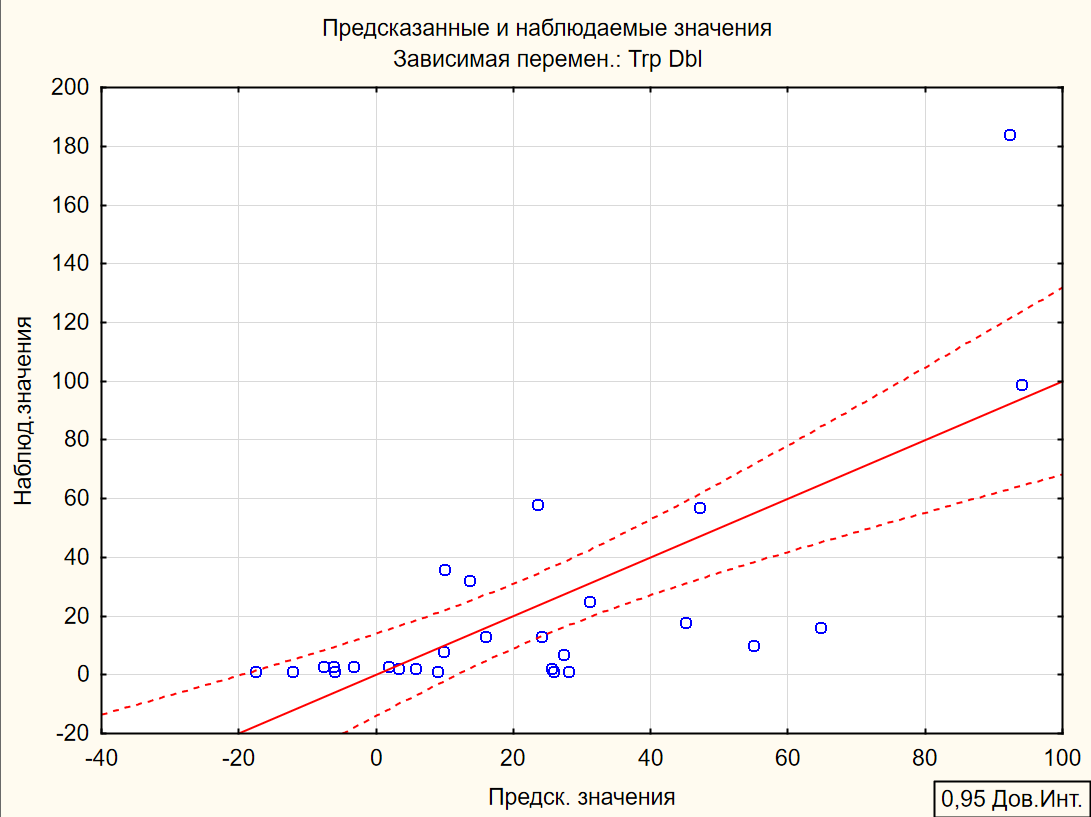
p-level говорит о значимости частных коэффициентов корреляции переменных AST и STL значимости p=0.1.

Изображение выглядит как стол

Автоматически созданное описание

Из таблицы видно, что статистика Дарбина-Уотсона имеет небольшое значение (2,15) при низкой сериальной корреляции (зависимость остатков) (-0,091).

Графическое сравнение предсказанных и наблюдаемых значений отклика приведено ниже:

****

****

По данной гистограмме можно судить о соответствии закона распределения остатков нормальному закону.

Можно сделать вывод о нормальной адекватности построенной линейной модели зависимости переменной Trpl\_Dbl от остальных количественных предикторов.

**8 Нелинейное многомерное моделирование взаимосвязей**

**8.1 Fixed Nonlinear Regression**

Этот модуль реализует множественный линейный регрессионный анализ с линеаризованной моделью.

Предварительно, до включения в модель множественной регрессии, можно просмотреть корреляции предикторов и функции отклика:

Изображение выглядит как текст, кроссворд, внутренний

Автоматически созданное описание

Из таблицы видно, что Очки хорошо коррелируют с подборами и передачами.

Изображение выглядит как текст, кроссворд

Автоматически созданное описание

Как видно, коэффициент детерминации высокий (). В модель не были включены Очки (PTS), и квадрат Подборов (V26\*\*2).

Для проверки адекватности модели построим гистограмму остатков.



Из гистограммы явно видно, что распределение остатков соответствует нормальному распределению.

**8.2 Логит регрессия**

Для построения Logit модели, добавим бинарную переменную NBA\_Champ, которая будет определять является ли игрок чемпионом NBA или нет.

Изображение выглядит как стол

Автоматически созданное описание

Построим Logit модель зависимости бинарной переменной NBA\_Champ от FG% (процент попаданий), 2P%, 3P% и FT% (процент штрафных попаданий).

Изображение выглядит как текст, кроссворд

Автоматически созданное описание

Регрессия незначима, так как p>0.05 (подтверждается гипотеза об отсутствии связи).

Нормальный вероятностный график остатков. Чем точки ближе к прямой, тем лучше.

Наблюдаемые и предсказанные

Изображение выглядит как стол

Автоматически созданное описание

**8.3 Пробит регрессия**

Построим Пробит модель зависимости бинарной переменной NBA\_Champ от FG% (процент попаданий), 2P%, 3P% и FT% (процент штрафных попаданий).

Изображение выглядит как текст, стол

Автоматически созданное описание

Регрессия незначима, так как p>0.05 (подтверждается гипотеза об отсутствии связи).

Нормальный вероятностный график остатков. Чем точки ближе к прямой, тем лучше.



Наблюдаемые и предсказанные:

Изображение выглядит как стол

Автоматически созданное описание

**8.4 Экспоненциальная регрессия**

Построим экспоненциальную модель зависимости Age(возраст) от PTS (кол-ва набранных очков)

Изображение выглядит как текст, стол

Автоматически созданное описание

Таблица предсказанных и наблюдаемых значений

Изображение выглядит как стол

Автоматически созданное описание

Гистограмма остатков:



Видно, что распределение остатков соответствует нормальному.

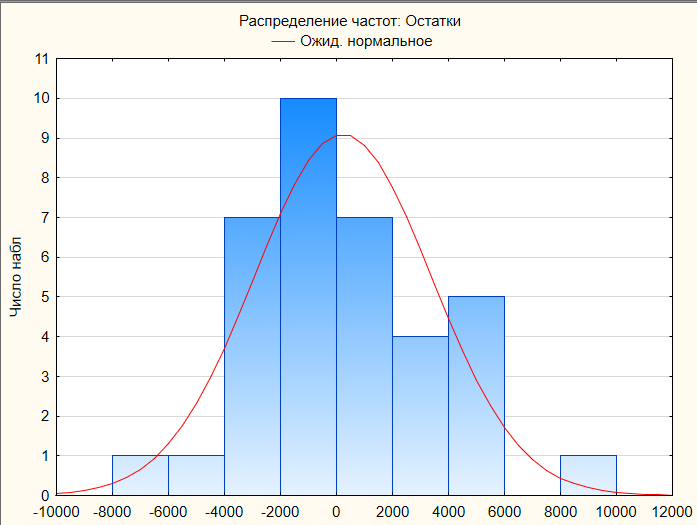
**8.5 Кусочно-линейная регрессия**

Построим кусочно-линейную модель зависимости Age(возраст) от PTS (кол-ва набранных очков).

Изображение выглядит как текст, стол

Автоматически созданное описание

Из таблицы видно, что доля объясненной дисперсии большая (80%), точка разрыва, определенная программой самостоятельно 12436,89



Из гистограммы видно, что распределение остатков соответствует нормальному распределению.

Таблица предсказанных и наблюдаемых значений:

Изображение выглядит как стол

Автоматически созданное описание

**8.6 Определенная пользователем регрессия**

Построим следующую модель: 3P%=FG% \* b0

Изображение выглядит как стол

Автоматически созданное описание

Из таблицы видно, что параметр значим.

Таким образом регрессия имеет вид:

3P%=FG% \* 0.735424

Нормальный вероятностный график остатков. Чем точки ближе к прямой, тем лучше.



Изображение выглядит как стол

Автоматически созданное описание

Из данной таблицы видно, что в целом регрессия значима.

**10 Дискриминантный анализ**

Цель дискриминантного анализа – на основе измерения различных характеристик объекта классифицировать его, то есть отнести к одной из нескольких групп некоторым оптимальным способом.

Задача состоит в том, чтобы по результатам рассмотрения параметров PTS, AST, TRB, STL отнести игроков по различным стилям игры (позициям).

Изображение выглядит как стол

Автоматически созданное описание

Лямбда Уилкса является результатом исключения соответствующей переменной из модели. Чем больше его значение, тем более желательно присутствие переменной в процедуре дискриминации. Частная Лямбда показывает отношение лямбды Уилкса после добавления соответствующей переменной к лямбде Уилкса. P-уровень – показатель уровня значимости критерия Фишера, если p<0.01 значит статистически значимый.

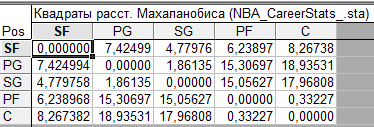


Таблица показывает расстояние между группами, по данной таблице можно судить о качестве дискриминантного анализа и степени различий (неоднородности) групп.

Изображение выглядит как стол

Автоматически созданное описание

Таблица показывает правильность оценивания и разбития данных на кластеры. Строки – наблюдаемые (исходные) классы. Столбцы – предсказанные классы.

Изображение выглядит как текст, стена, внутренний, кроссворд

Автоматически созданное описание

Проведя канонический анализ, мы получили 4 канонических корня для дискриминантных функций. Где первая строка дает критерий значимости для всех корней, вторая строка содержит значимость корней, оставшихся после удаления первого и тд. Таким образом таблица позволит оценить сколько значимых корней необходимо выделить. Как видно из таблицы, статистически значим только первый корень.

Изображение выглядит как текст, кроссворд, шкафчик, табло

Автоматически созданное описание

Изображение выглядит как текст, внутренний, кроссворд

Автоматически созданное описание

В первой таблице приведены стандартизованные переменные, по которым можно определить вклад переменных в каждую каноническую функцию. Из таблицы видно, что наибольший вклад в первую кан. Функцию вносит TRB, а во вторую AST.

Во второй таблице приведены внутригрупповые корреляции переменных с соответствующими дискриминантными функциями, их ещё называют структурными коэффициентами. Обычно используются для содержательной интерпретации функций. Из таблицы видно, что больше всего с первой функцией коррелирует TRB, а вот со второй PTS.

Изображение выглядит как текст, кроссворд, шкафчик, внутренний

Автоматически созданное описание

Таблица показывает средние значений для дискриминантных функций, которые лучше всего позволяют идентифицировать группы. Из таблицы видно, что первая функция в основном идентифицирует группу C и PF (игроков краски), вторая SF и тд.

Изображение выглядит как стол

Автоматически созданное описание

Таблица отражает значение каждой канонической функции для каждого наблюдения. Чем больше значение, тем лучше наблюдение относятся к той или иной функции.

Изображение выглядит как текст, кроссворд, стена, внутренний

Автоматически созданное описание

Таблица отображает функцию классификации, по которым наблюдения каждой группы могут быть классифицированы. Наблюдение приписывают той или иной группе, для которой функция имеет наибольшее значение

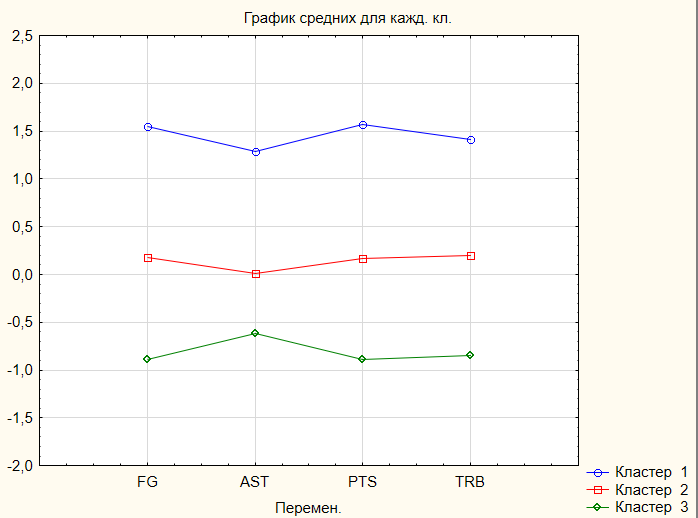
**11 Классификационный анализ без обучения**

**11.1 Кластерный анализ**

**11.1.1 Метод к-средних**

Кластерный анализ позволяет разбить множество исследуемых объектов и признаков на однородные в некотором смысле группы, или кластеры.

Попробуем разбить игроков на кластеры, но до этого надо стандартизировать переменные. Варьируя количеством кластеров и исключая переменные, получили следующие результаты:



Изображение выглядит как текст, внутренний, кроссворд

Автоматически созданное описание

Таблица и график свидетельствуют об успешной классификации игроков по параметрам FG, AST, PTS, TRB на три кластера. В таблицы приведены значения межгрупповых и внутригрупповых дисперсий признаков. Чем меньшее значение внутригрупповой и больше межгрупповой дисперсии, тем лучше признак характеризует принадлежность объектов к кластеру. Так как p<0.05 для всех признаков, то вклады всех признаков в разделение объектов на группы существенные.

Элементы кластеров и расстояние до центра кластера:

Изображение выглядит как стол

Автоматически созданное описание

Изображение выглядит как стол

Автоматически созданное описание

Изображение выглядит как стол

Автоматически созданное описание

Далее приведены средние для каждого кластера

Изображение выглядит как стол

Автоматически созданное описание

И Евклидовы расстояния между кластерами:

Изображение выглядит как стол

Автоматически созданное описание

**11.1.2 Двухвходовая кластеризация**

Проведем двухвходовую кластеризацию (одновременно и по переменным, и по строкам) для FG, AST, PTS, TRB.

Ниже приведен цветной график результата кластеризации:



Изображение выглядит как стол

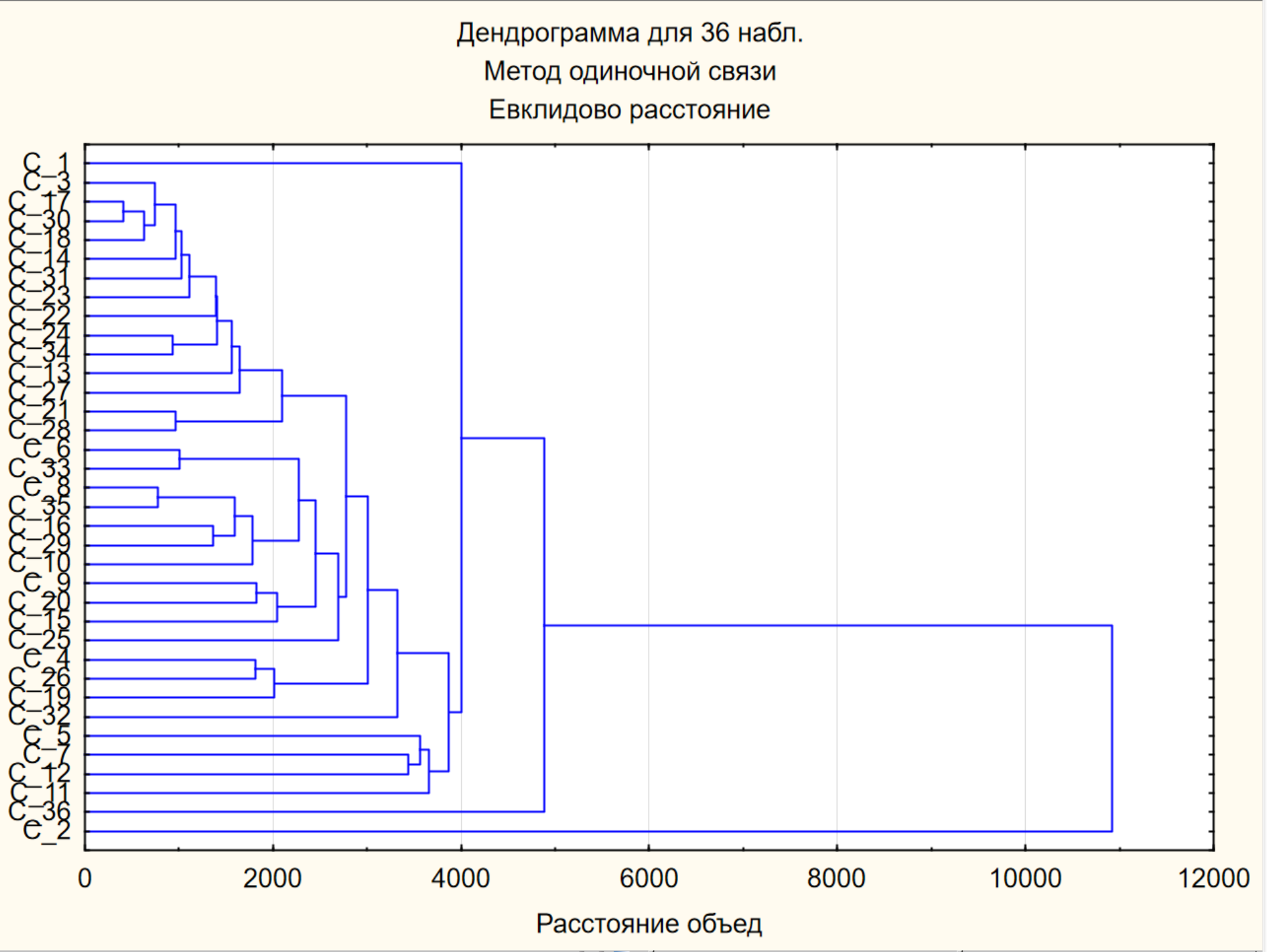
Автоматически созданное описание

В данной таблице приведены значения средних и стандартных отклонений для строк. (Для переменных аналогичные показатели равны 0 и 1 соответственно, т.к. переменные стандартизированы).

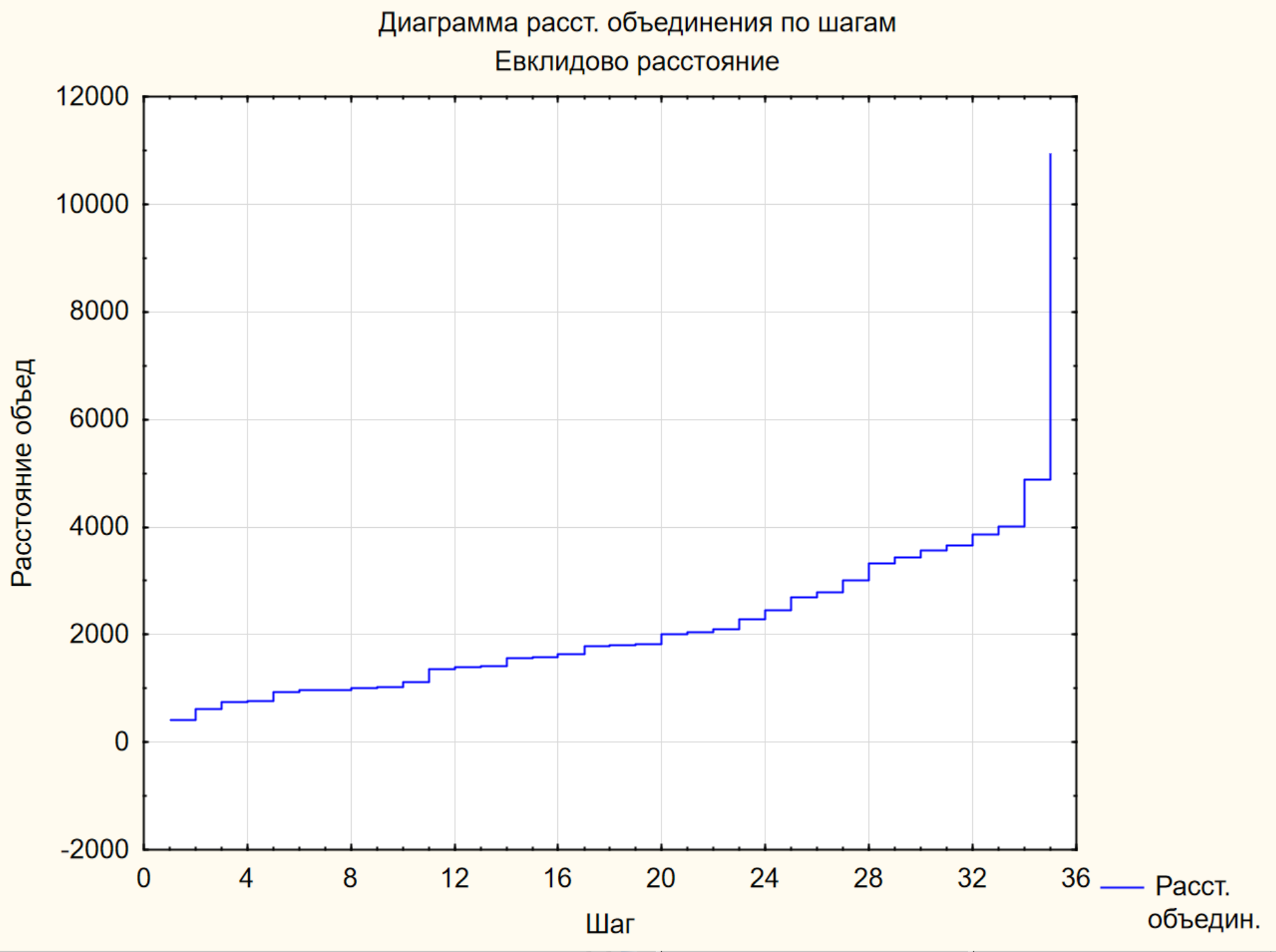
**11.1.3 Древовидная кластеризация**

Проведем древовиднуюкластеризацию по игрокам.

Горизонтальная древовидная диаграмма имеет вид:



На диаграмме видно, что по мере понижения порога (изменение расстояния объединения), относящегося к решению об объединении объектов в один кластер, все большее и большее число объектов связывается, и объединяются все больше кластеров.



На данном графике показан порядок объединения. Например, на 36-м шаге расстояние объединения должно было значительно вырасти, чтобы объединить еще переменные.

Фрагмент правила объединения в кластеры показано в следующей таблице:

Изображение выглядит как стол

Автоматически созданное описание

**11.2 Деревья классификации**

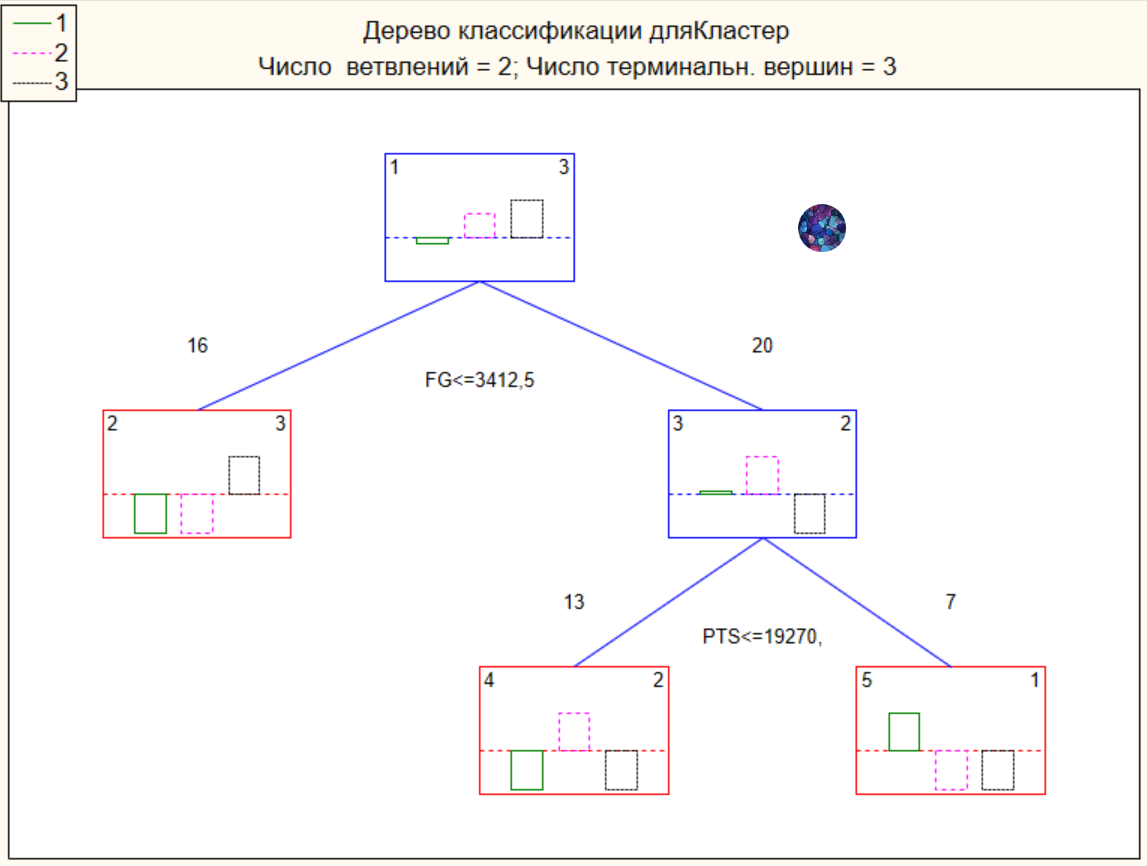
Деревья классификации – это метод, позволяющий предсказывать принадлежность наблюдений или объектов к тому или иному классу категориальной зависимой переменной в зависимости от соответствующих значений одной или нескольких независимых переменных.

Будим классифицировать игроков на кластеры (на мяче, с мячом и игрок краски), выявленные одним из методов кластеризации, описанным выше в зависимости от FG, TRB, AST, PTS.

Изображение выглядит как стол

Автоматически созданное описание

Граф дерева классификации изображен на следующем рисунке:



Условия ветвления:

Изображение выглядит как стол

Автоматически созданное описание

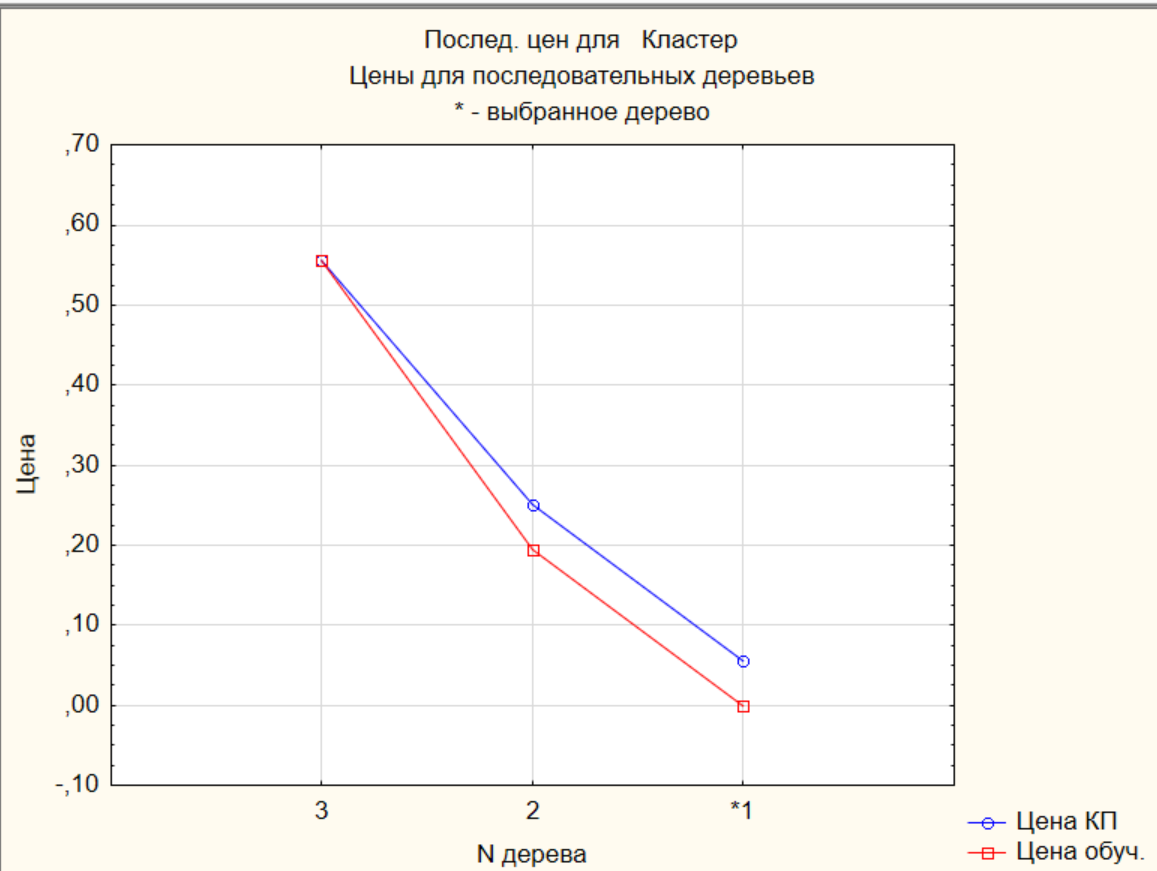
Последовательность деревьев:

Изображение выглядит как стол

Автоматически созданное описание

В таблице указаны терминальные вершины, цена кросс-проверки, ее стандартная ошибка, цена обучения и сложность каждого из усеченных деревьев. Звездой помечено дерево, которое было признано деревом «подходящего размера».

График последовательности цен :



Значимость предикторов:

Изображение выглядит как стол

Автоматически созданное описание

На рисунке показана дискретная карта линий уровня:

Изображение выглядит как квадрат

Автоматически созданное описание

Изображение выглядит как стол

Автоматически созданное описание

В этой таблице показаны априорные вероятности для каждого класса зависимой переменной и число элементов в каждом классе.

Ошибки классификации:

Изображение выглядит как стол

Автоматически созданное описание

В этой таблице показано, сколько объектов каждого класса было ошибочно отнесено к другому классу.

Кросс-проверка:

Изображение выглядит как стол

Автоматически созданное описание

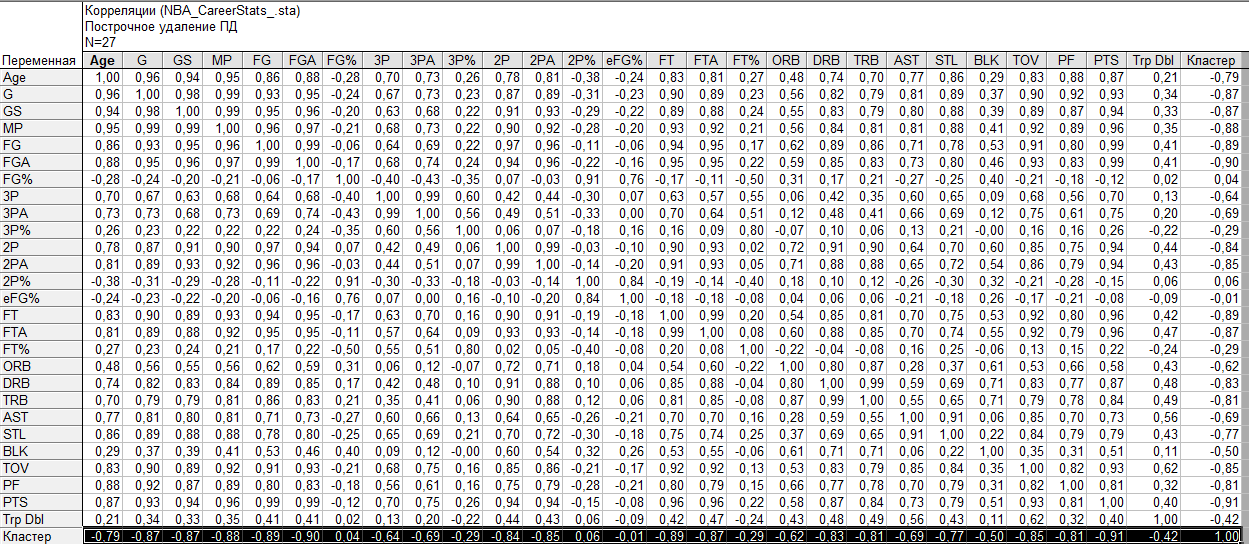
В данной таблице видно, что ошибки имеются при классификации второго класса, как третий при одном наблюдении. Кроме того, в таблице показана цена глобальной кросс-проверки и ее стандартное отклонение: 0,05556 и 0,03818 соответственно.

Таким образом, получено решающее правило, состоящее из двух этапов, которое позволит произвольного игрока классифицировать как на игроков на мяче, с мячом и игрока краски.

**12 Методы редукции данных**

**12.1 Факторный анализ**

Главная цель факторного анализа — это сокращение числа переменных и определение структуры взаимосвязей между переменными, т. е. классификация переменных.



Уберём все процентные переменные, Кластер и Age

Просмотрим собственные значения факторов:

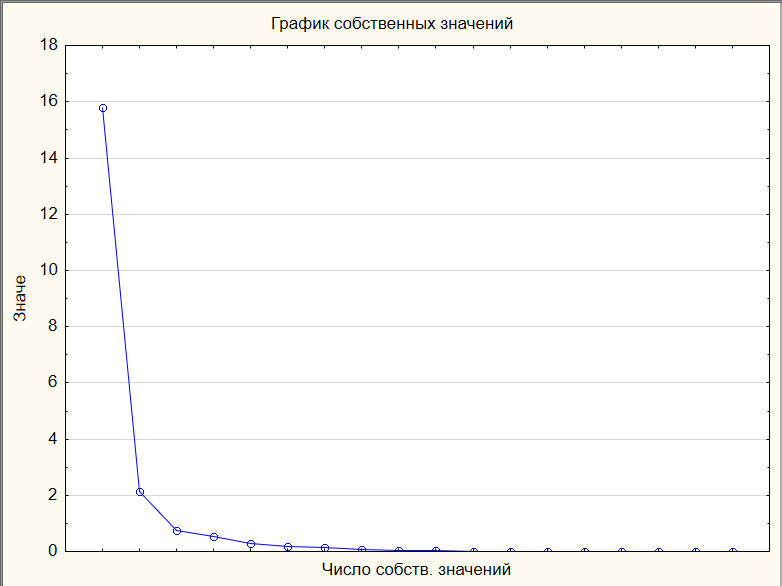
Изображение выглядит как стол

Автоматически созданное описание

Во втором столбце (Eigenvalue) приведены дисперсии выделенных факторов – собственные числа, в третьем – процент от общей дисперсии. Как видно, первый фактор объясняет 78,92% общей дисперсии, второй 10,65% и т. д. Необходимо понять, сколько факторов следует оставить.

По критерию Кайзера можем отобрать только факторы с собственными значениями, большими 1. Из таблицы видно, что на основе данного критерия выделяются только 2 фактора.

Критерий каменистой осыпи:



Надо выбрать такое место на графике, где убывание собственных значений слева направо максимально замедляется. Из графика видно, что в соответствии с этим критерием можно попытаться выделить 2 или 3 фактора.

Предположим, что число факторов неизвестно.

Корреляции между переменными и выделенными факторами:

Изображение выглядит как стол

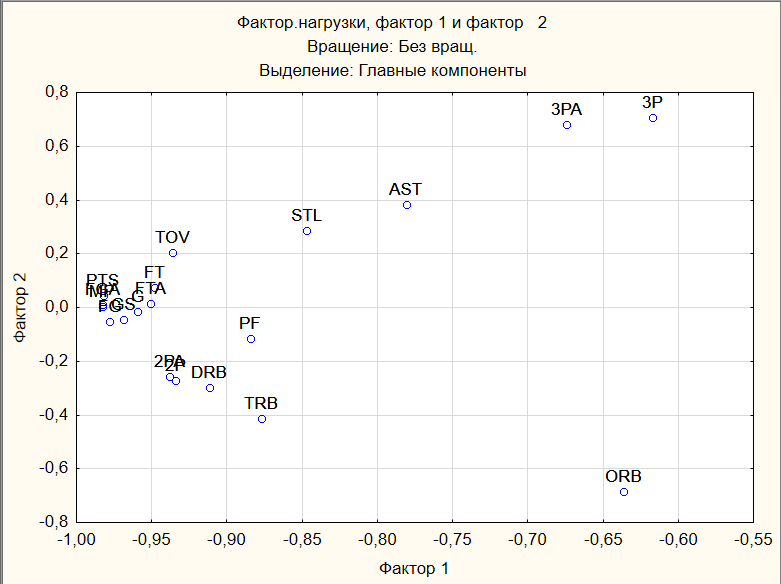
Автоматически созданное описание

Из таблицы видно, что первому и второму фактору соответствуют большие коэффициенты корреляции, чем остальным. Назначим число факторов 2.

Изображение выглядит как стол

Автоматически созданное описание

Из таблицы видно, что Factor 1 имеет высокие факторные нагрузки по переменным, в то время, когда Factor 2 имеет низкие факторные нагрузки по этим же переменным. Это означает, что выделенные два фактора наилучшим образом характеризуют данные.



Этот график иллюстрирует соотношение между факторами и группами переменных. Видно, что группа переменных AST, 3PA, 3P, ORB стоят отдельно от другой большой группы переменных.

Коэффициенты уравнений регрессий, по которым программа посчитает значения факторов для каждого наблюдения, представлены в таблице:

Изображение выглядит как стол

Автоматически созданное описание

Значения факторов для каждого игрока представлены в таблице:

Изображение выглядит как стол

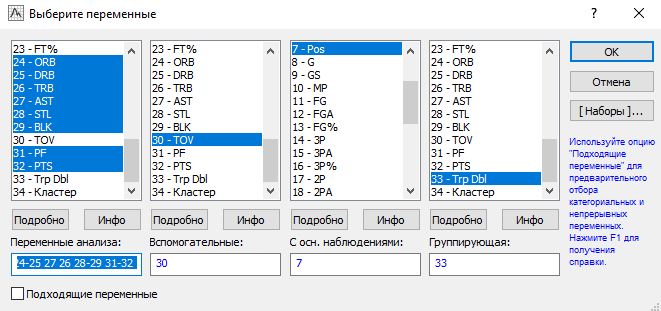
Автоматически созданное описание

По данным из этой таблицы можно судить об отношении игроков к соответствующему фактору. Положительное значение фактора соответствует позитивному отношению игрока, а отрицательное – негативному.

**12.2 Метод анализ главных компонент и классификация**

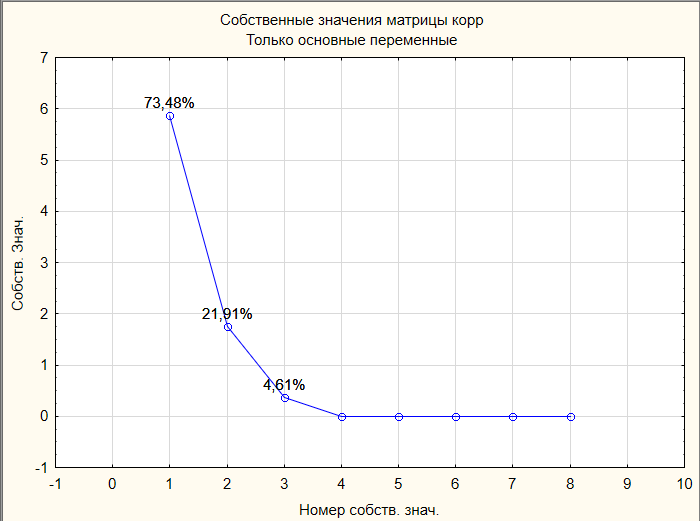
Метод анализ главных компонент и классификация служит для достижения двух целей: уменьшение общего числа переменных и классификация переменных и наблюдений, при помощи строящегося факторного пространства.

Для модуля были выбраны следующие параметры:



Проведём анализ для позиции PF.

График каменистой осыпи:



Из графика видно, что число выделяемых факторов может варьироваться от 2 до 3. Проверим этот факт по таблице собственных значений:

Изображение выглядит как текст, стол

Автоматически созданное описание

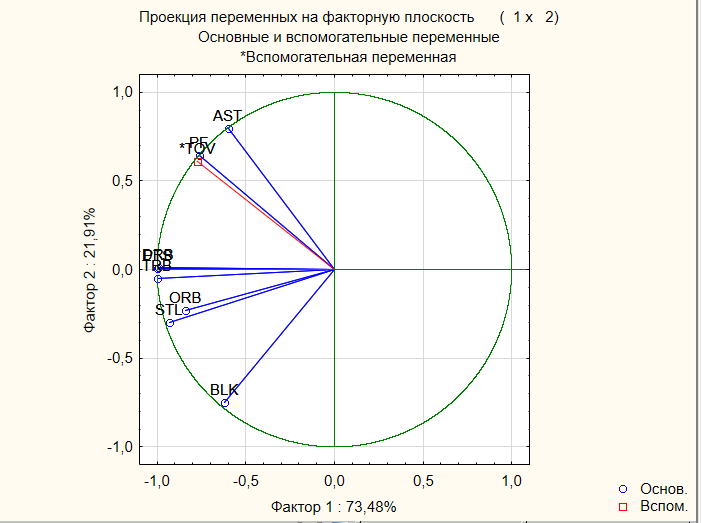
Эта таблица собственных значений. Во втором столбце (Eigenvalue) приведены дисперсии выделенных факторов – собственные числа, в третьем – процент от общей дисперсии. Как видно, первый фактор объясняет 73.5% общей дисперсии, второй – 21.9% и т. д. По критерию Кайзера можем отобрать только факторы с собственными значениями, большими 1. Из таблицы видно, что на основе данного критерия выделяются только 2 фактора.

Изображение выглядит как текст, кроссворд, с плиткой

Автоматически созданное описание

В данной таблице представлены факторные координаты переменных (факторные нагрузки). Большее абсолютное значение факторной нагрузки переменной с каким-либо фактором говорит о том, что переменная сильнее связана с этим фактором. Вспомогательные переменные обозначены «\*». Первая факторная ось, соответствующая собственному значению 5,88 наиболее сильно коррелирует с переменными PTS, DRB, TRB и STL. Вторая факторная ось, соответствующая собственному значению 1,75 коррелирует с AST, BLK, PF (персональные замечения).

График факторных координат переменных и наблюдений:



Чем ближе переменная к единичной окружности, тем лучше она воспроизведена в найденной системе координат (лучше воспроизводится текущим набором выделенных факторов).

Вклад переменных в дисперсию факторной оси:

Изображение выглядит как текст, кроссворд, с плиткой

Автоматически созданное описание

Изображение выглядит как стол

Автоматически созданное описание

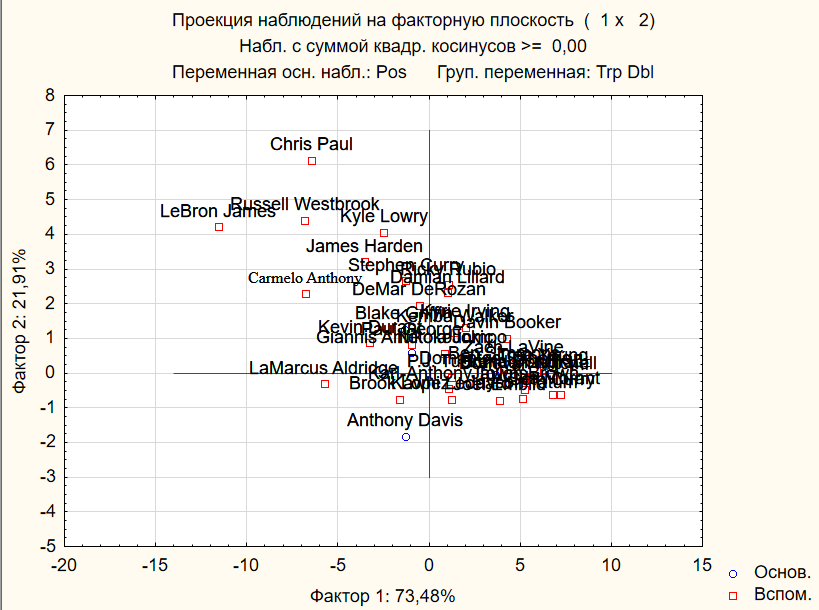
Это таблица общностей переменных. Общность – это доля объясненной дисперсии, которая характеризует степень общности переменной с другими переменными по заданному числу факторов. Из таблицы общностей переменных видно, что самая высокая степень общности с другими переменными для первого фактора у переменной PTS, у второго фактора - у TRB.

Изображение выглядит как стол

Автоматически созданное описание

В данной таблице представлена информация о принадлежности наблюдения к основным или вспомогательным наблюдениям.

График наблюдений в факторном пространстве:



**13 Многомерное шкалирование**

Цель многомерного шкалирования – поиск и интерпретация латентных переменных, дающих возможность пользователю объяснить сходства между объектами, заданными точками в исходном пространстве признаков.

Перед работой надо создать матричный файл. Его можно создать, например, при помощи кластерного анализа. Проведя кластеризацию по переменным FG, TRB, AST, STL, PTS, получим матричный файл, который представлен ниже:

Изображение выглядит как стол

Автоматически созданное описание

В результате оценивания параметров получим:

Изображение выглядит как стол

Автоматически созданное описание

Из данной таблицы видно, что на 21-й итерации получили лучшую двухмерную конфигурацию. D-зв вычисляется с помощью процедуры, которая пытается воспроизвести порядок следования рангов расстояний в исходной матрице расстояний. D-кр вычисляется с помощью метода, который позволяет подобрать монотонное преобразование (регрессию), наиболее точно воспроизводящее исходные расстояния.

Изображение выглядит как стол

Автоматически созданное описание

В этой таблице содержаться воспроизведенные расстояния и соответствующие им значения D-зв и D-кр. Для обозначения элементов используют матричный индекс, например, D(19,18) – 19 это строка матрицы входов, 18-колонкой.

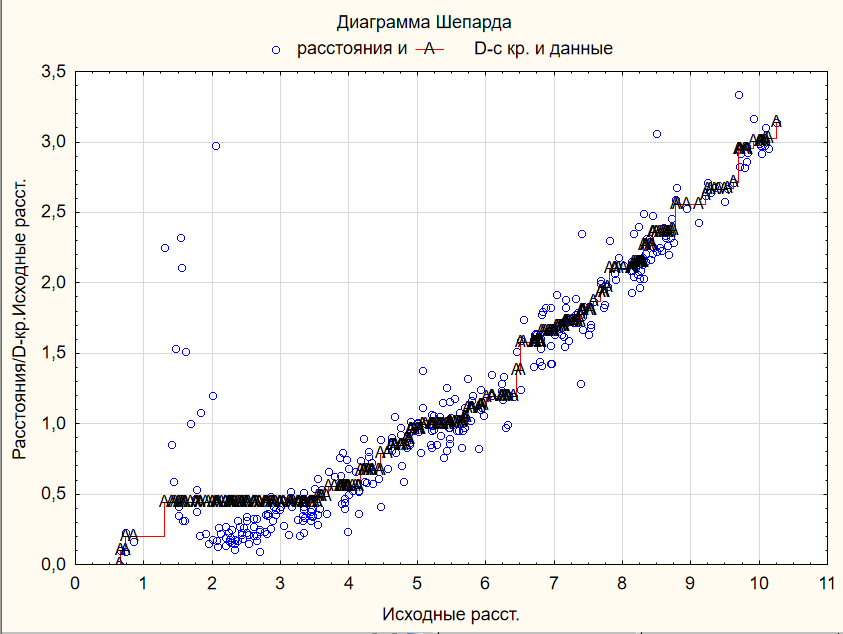
Изображение выглядит как стол

Автоматически созданное описание

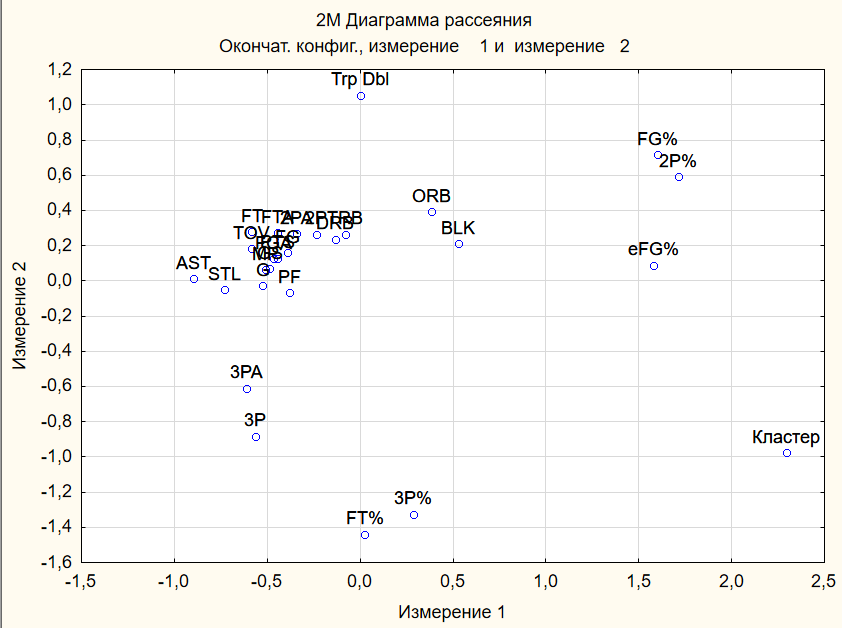
Таблица результатов с преобразованными входными значениями, вычисленными с помощью монотонной регрессии:

Изображение выглядит как текст

Автоматически созданное описание

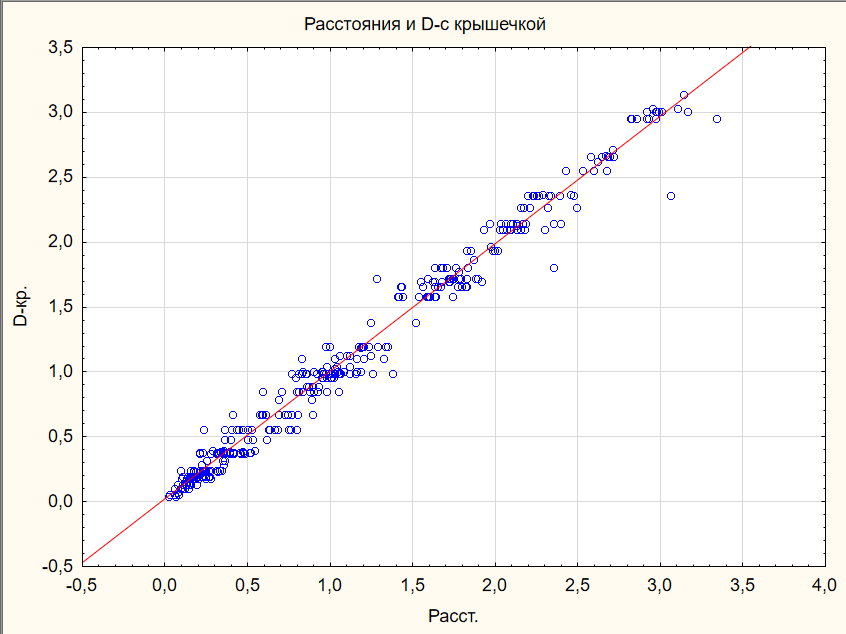


На этой диаграмме изображена зависимость воспроизведенных расстояний от исходных расстояний. Чем лучше согласие ступенчатой функции с точками данных на диаграмме рассеяния, тем лучше согласие с моделью (воспроизведение расстояний).



Это график окончательной конфигурации объектов на плоскости. Из графика видно, что Кластер стоит отдельно от остальной группы параметров, так же как и Trp\_Dbl, FT%, 3P% и тп.

График зависимости преобразованных значений входных данных от преобразованных расстояний:



Чем плотнее точки вокруг диагональной линии, тем лучше согласие с данными для выбранной модели.

График зависимости преобразованных значений входных данных от преобразованных расстояний:

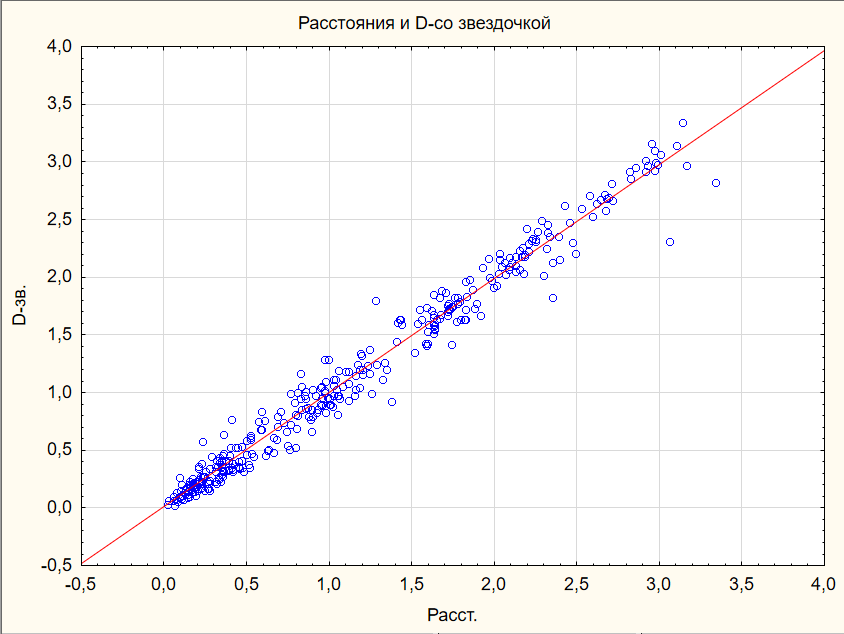
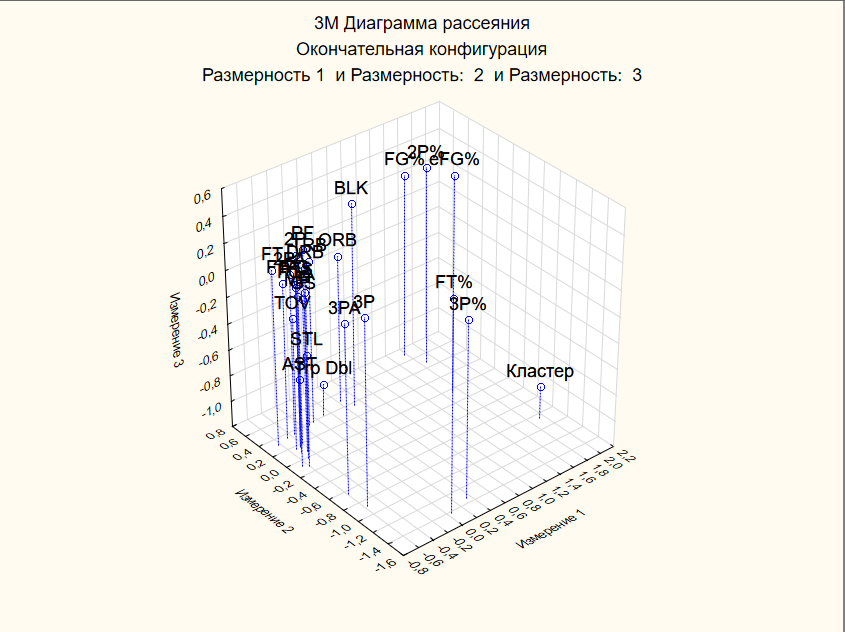


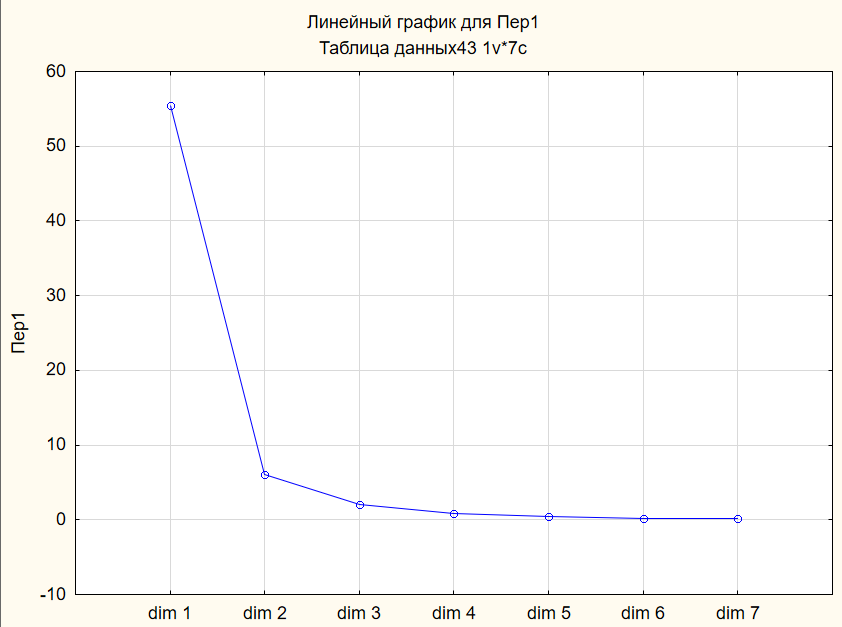
График окончательной конфигурации объектов в трехмерном пространстве.

Если выбрать размерность 3, то можем просмотреть график окончательной конфигурации объектов в трехмерном пространстве:



Для проверки правильности выбора размерности пространства воспользуемся критерием каменистой осыпи. Для этого проведем последовательно вычисления D-зв: raw stress для размерностей от 7 до 1, запишем полученные данные в таблицу и построим линейный график по полученной таблице.

|  |  |
| --- | --- |
|  | Значение параметра D-зв |
| Размерность 1 | 55.52701 |
| Размерность 2 | 6.023430 |
| Размерность 3 | 2.060916 |
| Размерность 4 | 0.9026368 |
| Размерность 5 | 0.4566985 |
| Размерность 6 | 0.2205953 |
| Размерность 7 | 0,1667798 |



Из графика согласно критерию каменистой осыпи следует, что для воспроизведения расстояний между страховщиками необходимо выбрать двухмерное пространство, так как в точке с абсциссой Dim 3 максимально замедляется уменьшение стресса. Следовательно, лучше всего проводить анализ в пространстве размерности 3.