**Ники Георгиев Манавски, СИ III курс, фак.№ 1901321083**

**КУРСОВА ЗАДАЧА:**

Направете 85% случайна извадка от файла RED\_winе\_2.sav. Запишете данните в нов файл с име <ФNo\_фамилия>. Файлът съдържа информация за химическия състав на марка червено вино както следва:

Quality – качество (средно аритметична оценка от 1 до 10)

Fixed acidity – Фиксирана киселинност (g/dm3);

Volatile acidity – Летливи киселини (g/dm3);

Citric acid – Лимонена киселина (g/dm3);

Residual sugar – Остатъчна захар (g/dm3);

Chlorides – Хлориди (g/dm3);

Free sulfur dioxide – Свободен серен диоксид (mg/dm3);

Total sulfur dioxide – Общ серен диоксид (mg/dm3);

Density – Плътност (g/cm3);

рН;

Sulphates – Сулфати (g/dm3);

Alcohol - Алкохол (%);

1. /5 т./ Направете описателна статистика променливата ПЛЪТНОСТ на виното. Запишете мерките за централна тенденция и мерките за разсейване. Тълкувайте значението на коефициентите от гледна точка на нормалност на разпределението.

2. /5 т./ Начертайте хистограма, бокс-плот и Q-Q plot на изследваната променлива. Тълкувайте графиките от гледна точка на нормалност на разпределението.

3. /5 т./ Направете тестове за нормалност и установете вида на разпределението на променливата ПЛЪТНОСТ на виното.

4. /5 т./ Проверете хипотезата, че средната плътност на виното е 0,9 срещу алтернативната двустранна хипотеза.

5. /5 т./ Установете има ли статистическа значимо различие в плътността на виното при оценка за качеството му под и над 7,5.

6. /5 т./ Изследвайте със статистически анализи наличието на възможни зависимости на плътността на виното от останалите променливи. Запишете всички корелационни коефициенти и тълкувайте техните знаци, стойности и тяхната значимост. Определете факторите, от които плътността се влияе най-силно.

7. /5 т./ Постройте адекватен многомерен линеен регресионен модел за плътността на виното, като за предиктори изберете всички останали променливи. Опишете регресионния модел – коеф. на детерминация, адекватност на модела, значимост на коефициентите и др. Интерпретирайте получените резултати.

8. /5 т./ Постройте нов многомерен линеен регресионен модел за плътността на виното, като отстраните незначимите предиктори, установени в модела от т.7, ако има такива. Опишете показателите на новия регресионен модел. Коментирайте го – подобриха ли се?

9. /5 т./ Запазете предсказаните данни от модела и запишете уравнението на регресия, което е с по-добри показатели.

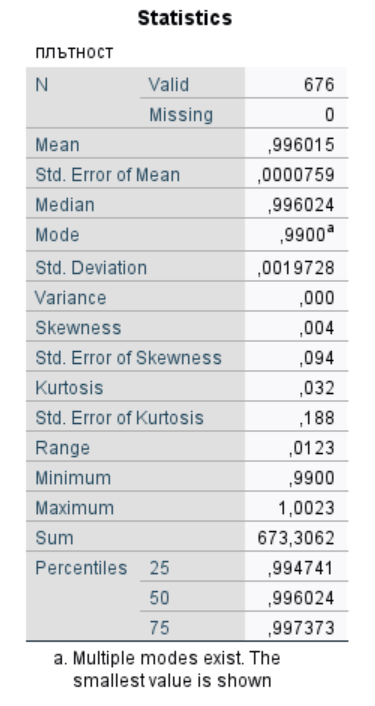
10. /5 т./ Изследвайте разпределението на остатъците с тестове и графики. Направете изводи.

11. /5 т./ Постройте CART модел на плътността на виното и останалите предиктори с минимум 100 случая в родителски и 50 в наследствен възел. Определете кои предиктори имат най-силно влияние.

12. /5 т./ Отстранете предикторите с минимално влияние и постройте нов CART модел. Запишете предсказаните от него стойности. Подобри ли се коефициента на детерминация на новия CART модел в сравнение с регресионния?

**Решение:**

1. **Статистика за плътност**

****

**Обем на извадката:** 676

**Мерки за централна тенденция:**

* ***Медиана:*** ,996024
* ***Мода:*** ,9900^a
* ***Средно аритметична стойност:*** ,996015

**Мерки за разсейване:**

* ***Размах:*** ,0123
* ***Дисперсия:*** ,000
* ***Стандартно отклонение:*** ,0019728

**Квартили:**

**Q1:** ,994741 **Q2:** ,996024 **Q3:** ,997373

**Skewness:** ,004

**Kurtosis:** ,032

**p=1,96**

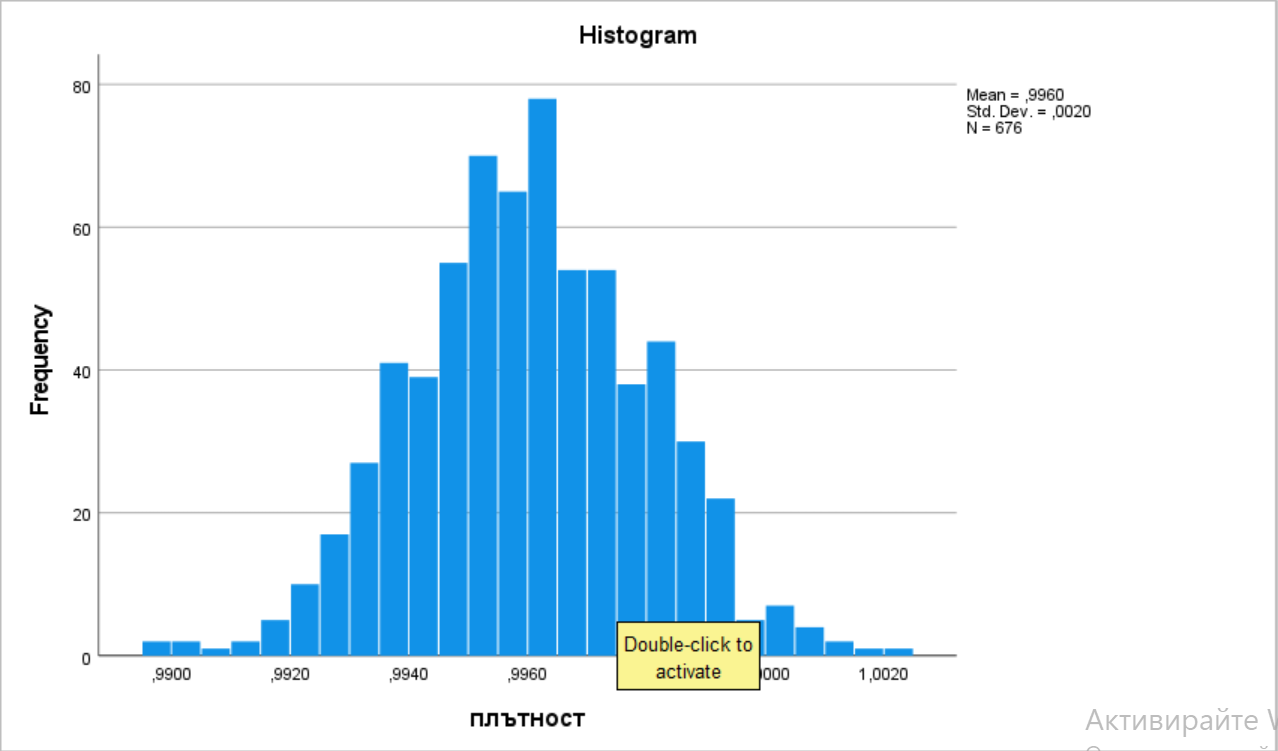
**|Sk|/SE(Sk) < p => 0,004/0,094=0,0425<p**

**|Ku|/SE(Ku) < p => 0,032/0,188=0,1702<p**

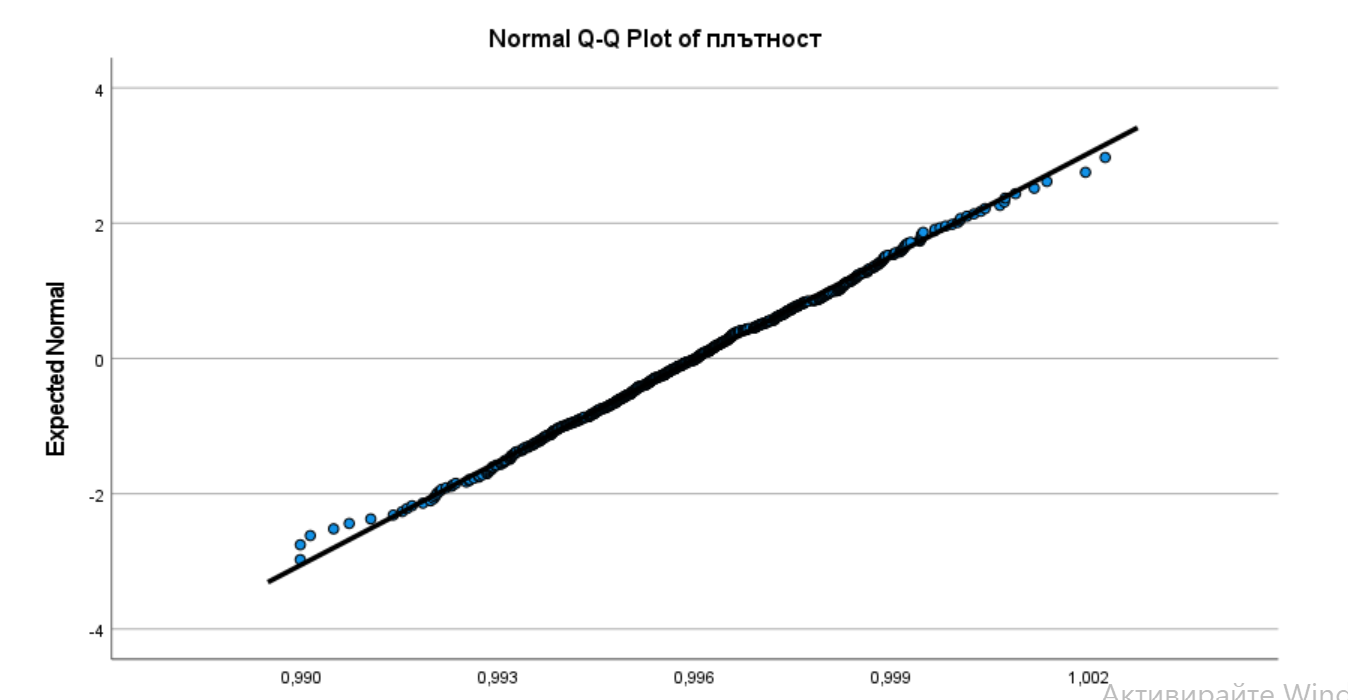
* **са изпълнени неравенствата => е нормално разпределена**

**Данните са нормално разпределени, тъй като средната стойност, медианата и модата са приблизително равни, но се наблюдава дясна асиметрия(струпване на данните в лявата страна на извадката) и опашките на разпределението са с голяма тежест.**

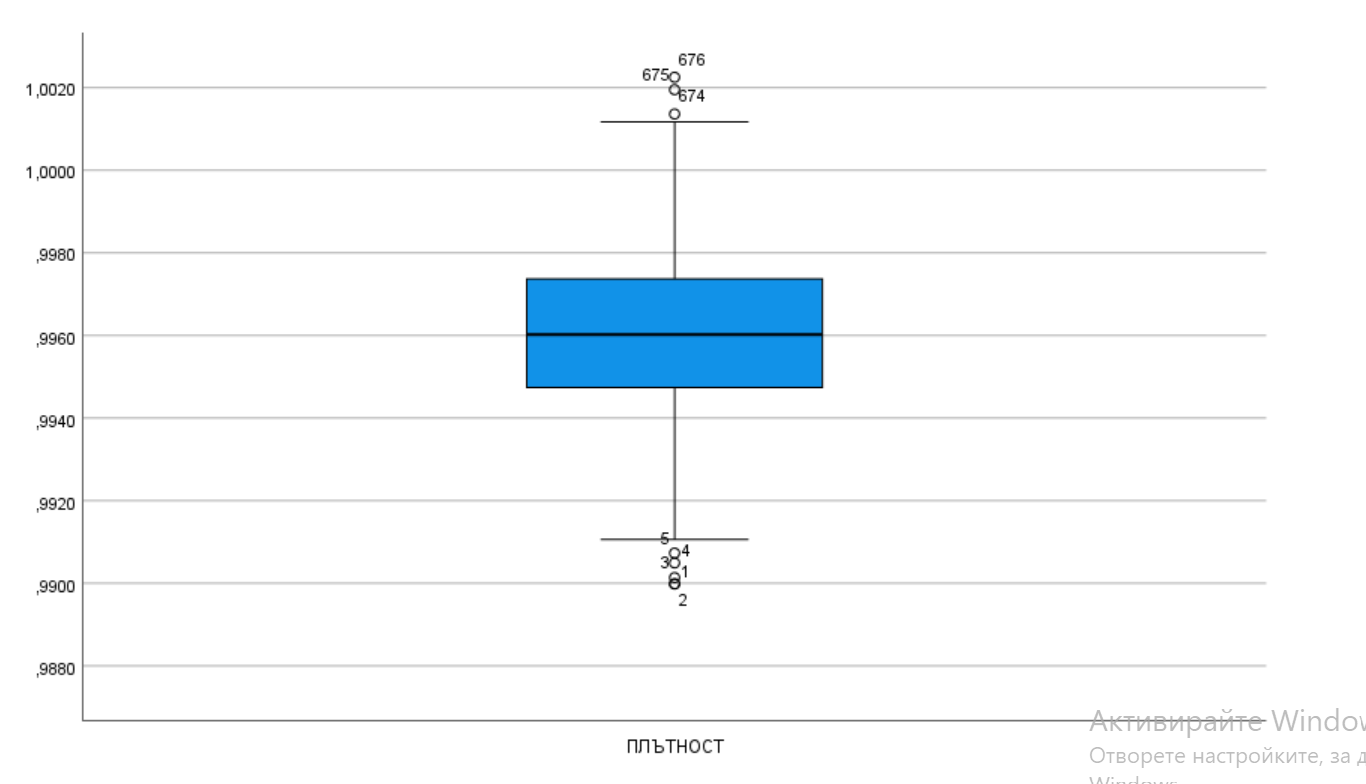
1. **Хистограма, бокс-плот и Q-Q Plot**

****

**От хистограмата можем да видим, че разпределението на извадката не е съвсем нормално(има интервали, чиято стойност не достига кривата на нормалното разпределение)**

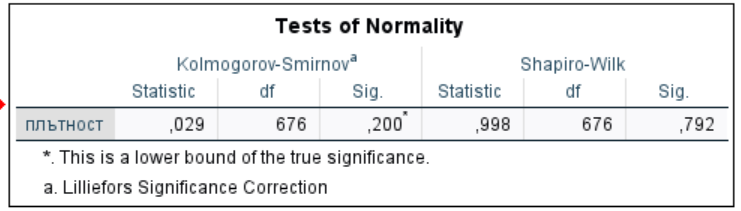
****

**Разпределението на извадката е нормално, но не всички стойности се доближават до правата(има леко разпръсване в краищата).**

****

**Разпределението на извадката е нормално, но има стойности, които са извън мустаците на бокс-плота.**

1. **Тестове за нормалност**

****

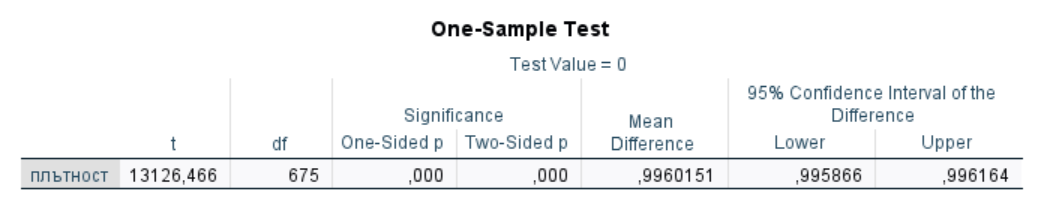
**H0: Разпределението е нормално**

**H1: Разпределението не е нормално**

**Тест на Kolmogorov-Smirnoff: Тест на Shapiro-Wilk:**

Sig. = ,200 > 0,05 => H0 се приемаSig. = ,792 > 0,05 => H0 се приема

1. **Хипотеза за средната плътност**

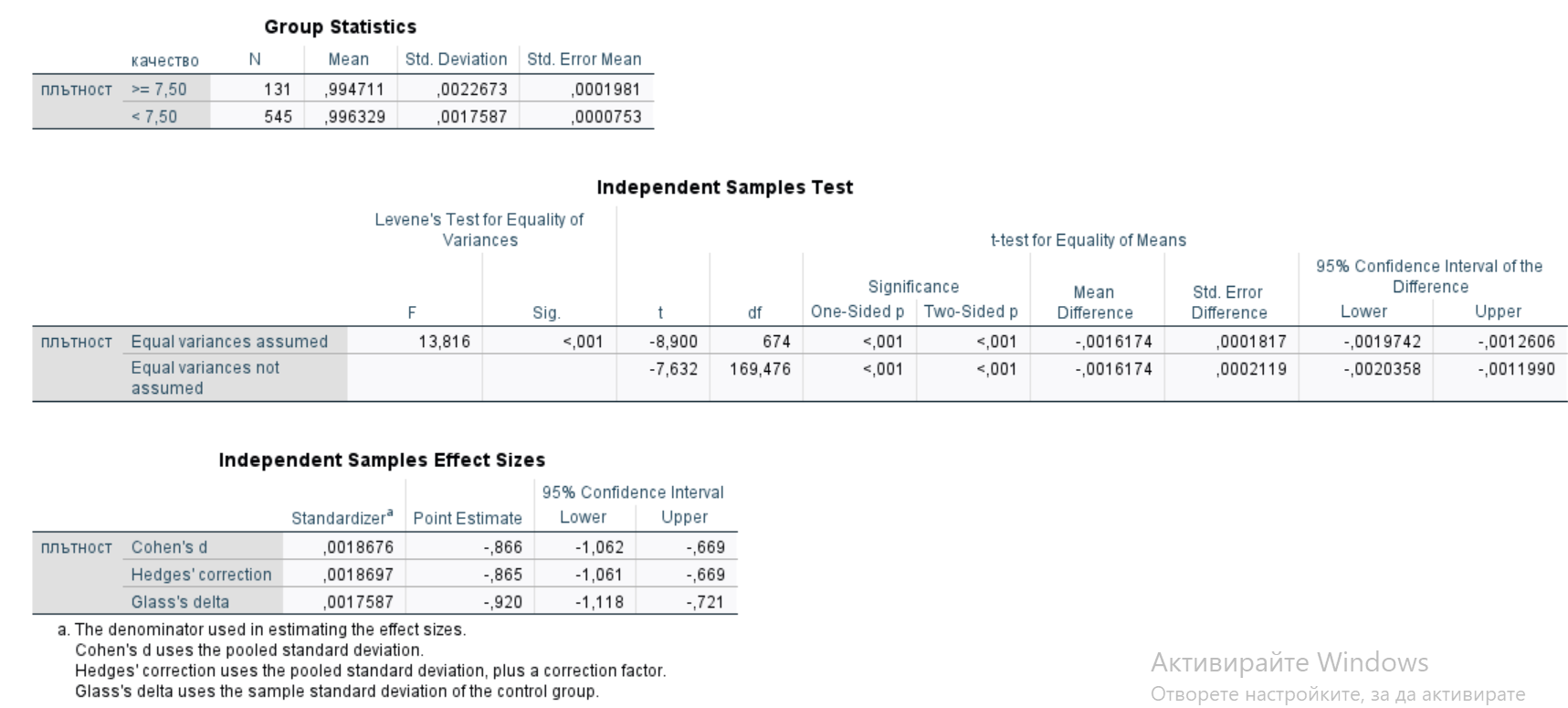
****

**H0: Средната плътност на виното е 0,9**

**H1: Средната плътност на виното не е 0,9**

**Sig. e ,000 < 0,05 => отхвърляме H0**

1. **Има ли статистическа значимо различие в плътността на виното при оценка за качеството му под и над 7,5?**

****

**Правим проверка дали има статистическо различие в плътността на виното спрямо неговото качество(под и над 7,5)**

**T-Test за 2 независими извадки**

**H0:**  σ1=σ2

**H1:**  σ1/=σ2

**Sig. e под 0,001, което е < 0,05 => отхвърляме H0 и приемаме H1**