**INTRODUCERE**

Analiza datelor este un proces complex care presupune integrarea diferitelor metodologii si tehnici pentru a obține informații valoroase despre un set de date. Aceasta poate fi folosită pentru a identifica tendințe, modele, fapte și concluzii care pot fi utilizate pentru a lua decizii informate sau pentru a îmbunătăți înțelegerea unui fenomen.

În acest proiect, vom utiliza analiza datelor pentru a examina caracteristicile vinului alb și vinul roșu, cum ar fi aciditatea fixă, aciditatea volatilă, aciditatea citrică, zahărul rezidual, clorura, dioxidul de sulf liber, dioxidul de sulf total, densitatea, pH-ul, sulfații, alcoolul și calitatea. Scopul acestui proiect este de a înțelege relația dintre aceste caracteristici și calitatea vinului, precum și de a identifica eventualele diferențe între vinul alb și vinul roșu.

Pentru a atinge acest scop, vom utiliza diverse tehnici de analiză de date, cum ar fi descrierea statistică, analiza exploratorie de date și modelarea. Descrierea statistică ne va oferi o imagine generală a datelor, inclusiv informații despre medie, deviația standard, etc. Analiza exploratorie de date ne va permite să identificăm tendințe, relații și modele în datele noastre. Modelarea va fi utilizată pentru a crea modele predictive care pot fi utilizate pentru a îmbunătăți înțelegerea noastră despre aceste produse și pentru a lua decizii informate.

În concluzie, analiza datelor este un proces esențial care ne poate ajuta să înțelegem mai bine fenomenele din jurul nostru și să luăm decizii informate. Procesul de analiză a datelor este continuu și presupune integrarea diferitelor tehnici și metodologii pentru a obține informații valoroase despre un set de date.

**Analiza univariata**

Fisierele privind datele despre vin rosu si vin alb se afla in fisiere tip CSV cu valorile delimitate prin “;”. Pentru a citi mai usor datele, am convertit fisierele in tip XLSX, urmarind urmatorii pasi:

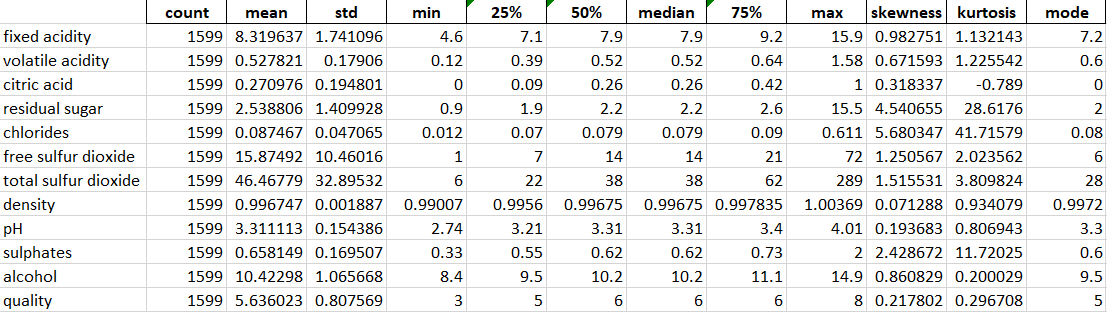
* Am deschis Microsoft Excel
* Am dat click pe File, apoi pe Open
* Am navigat la locația fișierului CSV și l-am selectat
* Am schimbat "Files of type" în "Text Files"
* Am selectat fișierul CSV și am facut click pe Open
* In wizard-ul de import text, am selectat "Delimitat" și "Punct și virgulă" ca delimitator.
* Am salvat fisierul ca XLSX, făcând click pe File, apoi Save As "Excel Workbook" ca format.

Urmatorul aspect in vederea analizei corecte a datelor a fost identificarea valorilor lipsa din fisier.

Folosind functia print(df.isnull().sum()), am obtinut valoarea 0 aferenta tuturor coloanelor, ceea ce indica faptul ca nu avem valori lipsa in baza de date privind vinul rosu.

Astfel, putem continua analiza, obtinand datele aferente statisticilor descriptive, incluzand si valorile pentru “skewness”, “kurtosis”, mediana si valoare modala.

Privind vinul rosu, am obtinut urmatoarele rezultate:



Tabel 1. Statistici descriptive pentru vin rosu

Aceste date vor fi interpretate pentru o intelegere mai buna a caracteristicilor fiecarei variabile.

Aciditate fixă: Medie: 8,32 (Nivelul mediu al acidității fixe este 8,32) Mediană: 7,9 (Valoarea din mijloc a nivelurilor de aciditate fixă este 7,9) Modă: 7,2 (Nivelul cel mai frecvent întâlnit al acidității fixe este 7,2) Asimetrie: 0,98 (Distribuție cu asimetrie pozitivă, indicând mai multe valori spre aciditate mai mare) Curtoza: 1,13 (Distribuție platikurtica, indicând o distribuție plata în comparație cu normalul)

Aciditate volatilă: Medie: 0,53 (Nivelul mediu al acidității volatile este 0,53) Mediană: 0,52 (Valoarea din mijloc a nivelurilor de aciditate volatilă este 0,52) Modă: 0,6 (Nivelul cel mai frecvent întâlnit al acidității volatile este 0,6) Asimetrie: 0,67 (Distribuție cu asimetrie pozitivă, indicând mai multe valori spre aciditate mai mare) Curtoza: 1,23 (Distribuție platikurtica, indicând o distribuție plata a datelor în comparație cu normalul)

Acid citric: Medie: 0,27 (Nivelul mediu al acidității citrice este 0,27) Mediană: 0,26 (Valoarea din mijloc a nivelurilor de acid citric este 0,26) Modă: 0 (Nivelul cel mai frecvent întâlnit al acidității citrice este 0) Asimetrie: 0,32 (Distribuție cu asimetrie pozitivă, indicând mai multe valori spre aciditate mai mare) Curtoza: -0,79 (Distribuție platocurtică, indicând o distribuție mai plată în comparație cu normalul)

Zahăr rezidual: Medie: 2,54 (Nivelul mediu al zahărului rezidual este 2,54) Mediana: 2,2 (Valoarea mediană a zahărului rezidual este 2,2) Minim: 1,6 (Valoarea minimă a zahărului rezidual este 1,6) Maxim: 3,5 (Valoarea maximă a zahărului rezidual este 3,5)

Cloruri: Media: 0.09 (Nivelul mediu al clorurilor este 0.09) Mediana: 0.079 (Valoarea mediană a nivelurilor clorurilor este 0.079) Moda: 0.08 (Nivelul cel mai frecvent întâlnit al clorurilor este 0.08) Asimetria: 5.68 (Distribuție pozitiv asimetrică, indicând mai multe valori spre niveluri mai mari ale clorurilor) Curtoza: 41.72 (Distribuție extrem de leptocurtică, indicând o distribuție cu multe valori mari comparativ cu normalul)

Dioxid de sulf liber: Media: 15.87 (Nivelul mediu al dioxidului de sulf liber 15.87) Mediana: 14 (Valoarea mediană a nivelurilor dioxidului de sulf liber este 14) Moda: 6 (Nivelul cel mai frecvent întâlnit al dioxidului de sulf liber este 6) Asimetria: 1.25 (Distribuție pozitiv asimetrică, indicând mai multe valori spre niveluri mai mari ale dioxidului de sulf) Curtoza: 2.02

Dioxid de sulf total: Media dioxidului de sulf total este 46.47 cu o deviație standard de 32.89. Valoarea mediană este 38, indicând că jumătate dintre observații au o valoare mai mică și jumătate au o valoare mai mare. Valoarea modală nu este furnizată în rezultat, indicând că probabil nu este unică. Asimetria este 1.52, indicând o asimetrie moderat pozitivă, însemnând că coada pe partea pozitivă a mediei este mai lungă decât coada pe partea negativă a mediei. Curtoza este 3.81, distributia fiind leptokurtica, indicând un nivel moderat de înălțime.

Densitate: Densitatea medie este 0,997 cu o deviatie standard de 0,002. Valoarea mediană este 0,997, indicând că jumătate din observații au o valoare mai mică și jumătate au o valoare mai mare. Valoarea modală nu este furnizată în rezultat, indicând faptul că probabil nu este unică. Skewness este 0,071, indicând o ușoară înclinare pozitivă, însemnând că coada de pe partea pozitivă a mediei este mai lungă decât coada de pe partea negativă a mediei. Kurtosis este 0,934, indicând o distribuție relativ plată.

pH: pH-ul mediu este 3,31 cu o deviație standard de 0,154. Valoarea mediană este 3,31, indicând că jumătate din observații au o valoare mai mică și jumătate au o valoare mai mare. Valoarea modală nu este furnizată în rezultat, indicând faptul că probabil nu este unică. Skewness este 0,194, indicând o ușoară înclinare pozitivă, însemnând că coada de pe partea pozitivă a mediei este mai lungă decât coada de pe partea negativă a mediei. Kurtosis este 0,807, indicând o distribuție relativ plată.

Sulfați: Cantitatea medie de sulfați din mostrele de vin este de 0,658 g/dm^3. Cantitatea mediana este de asemenea 0,62 g/dm^3. Valoarea modală este 0,6 g/dm^3. Valoarea skewness este 2,428, ceea ce sugerează că distribuția sulfaților are o înclinare pozitivă. Acest lucru înseamnă că există mai multe mostre cu valori mai mari decât valori mai mici. Valoarea kurtosis este 11,72, ceea ce indică un vârf înalt al distribuției și prezența outlierilor.

Alcool: Continutul mediu de alcool al mostrelor de vin este de 10,42%. Continutul mediu de alcool este de 10,2%. Valoarea modala este de 9,5%. Valoarea asimetriei este 0,86, sugerand o usoara asimetrie pozitiva. Valoarea curtozei de 0,20 este relativ scazuta, indicand o distributie platikurtica.

Calitatea: Scorul mediu de calitate al mostrelor de vin este de 5,64, pe o scala de la 0 la 10. Scorul mediu este de 6. Valoarea modala este 5. Valoarea asimetriei este 0,22, ceea ce sugereaza o usoara asimetrie pozitiva, insemnand ca sunt mai multe mostre cu scoruri mai mari de calitate. Valoarea curtozei de 0,30 este relativ scazuta, indicand o distributie relativ plat.

Privind vinul alb, am obtinut urmatoarele rezultate:  


Tabel 2. Statistici descriptive pentru vin alb

Idem ca pentru statisticile privind vinul rosu, vom efectua o interpretare sumara a datelor, pentru o claritate sporita privind distributia acestora.

Fixed Acidity: Valoarea mediana a acizilor fixi din vin este de 6.8, cu o valoare modala identica de 6.8. Skewness-ul este de 0.65, indicand o usoara tendinta spre asimetrie pozitiva. Kurtosis-ul este de 2.17, indicand o distributie mai aplatizata decat distributia normala.

Volatile Acidity: Valoarea mediana a acizilor volatili din vin este de 0.26, cu o valoare modala de 0.28. Skewness-ul este de 1.58, indicand o asimetrie pozitiva puternica. Kurtosis-ul este de 5.09, indicand o distributie foarte ascutita.

Citric Acid: Valoarea mediana a acidului citric din vin este de 0.32, cu o valoare modala de 0.3. Skewness-ul este de 1.28, indicand o asimetrie pozitiva moderata. Kurtosis-ul este de 6.17, indicand o distributie foarte ascutita.

Residual Sugar: Valoarea mediana a zaharului rezidual din vin este de 5.2, cu o valoare modala de 1.2. Skewness-ul este de 1.08, indicand o asimetrie pozitiva moderata. Kurtosis-ul este de 3.47, indicand o distributie mai ascutita decat distributia normala.

Chlorides: Valoarea mediana a clorurilor din vin este de 0.043, cu o valoare modala identica de 0.044. Skewness-ul este de 5.02, indicand o asimetrie pozitiva puternica. Kurtosis-ul este de 37.57, indicand o distributie foarte aplatizata.

Free Sulfur Dioxide: Valoarea mediana a dioxidului de sulf liber din vin este de 34, cu o valoare modala de 29. Skewness-ul este de 1.41, indicand o asimetrie pozitiva moderata. Kurtosis-ul este de 11.47, indicand o distributie mai foarte ascutita fata de o distributia normala.

Total Sulfur Dioxide: Valoarea mediana a totalului de dioxid de sulf din vin este de 134, cu o valoare modala de 111. Skewness-ul este de 0.39, indicand o usoara tendinta spre asimetrie negativa. Kurtosis-ul este de 0.57, indicand o distributie aplatizata.

Density: Densitatea are o medie de 0.994027376, o valoare modală de 0.99374, un skewness de 0.977773005 si un kurtosis de 9.793806911. Valoarea medie indică că densitatea vinului este relativ constantă. Valoarea modală este aproape de aceeași valoare cu mediana (0.99374 vs 0.994027376), ceea ce indică că distribuția este simetrică. Skewness pozitiv indică că există o coadă mai lungă în partea dreaptă a distribuției, în timp ce kurtosis pozitiv indică că există o concentrație mai mare de valori în jurul mediei, decât într-o distribuție normală. În general, aceste valori indică o distribuție relativ normală a densității vinului.

pH: Valoarea mediana a pH-ului din setul de date este de 3.18, ceea ce inseamna ca jumatate din valorile pH sunt mai mari si jumatate sunt mai mici decat 3.18. Valoarea modala a pH-ului este de 3.14, aceasta fiind valoarea care apare cel mai des in setul de date. Skewness (abaterea de la simetrie) a pH-ului este de 0.46, ceea ce inseamna ca distributia valorilor pH nu este simetrica si poate fi inclinata intr-o anumita directie. Kurtosis (abaterea de la forma normala) a pH-ului este de 0.53, ceea ce inseamna ca distributia valorilor pH poate fi aplatizata sau incrancenata fata de o distributie normala.

Sulphates: Valoarea mediana a sulfatilor din setul de date este de 0.47, ceea ce inseamna ca jumatate din valorile sulfatilor sunt mai mari si jumatate sunt mai mici decat 0.47. Valoarea modala a sulfatilor este de 0.5, aceasta fiind valoarea care apare cel mai des in setul de date. Skewness a sulfatilor este de 0.98, ceea ce inseamna ca distributia valorilor sulfatilor nu este simetrica si poate fi inclinata intr-o anumita directie. Kurtosis a sulfatilor este de 1.59, ceea ce inseamna ca distributia valorilor sulfatilor poate fi mai aplatizata decat o distributie normala.

Alcohol: Valoarea mediana a alcoolului din setul de date este de 10.4, ceea ce inseamna ca jumatate din valorile alcoolului sunt mai mari si jumatate sunt mai mici decat 10.4. Valoarea modala a alcoolului este de 9.4, aceasta fiind valoarea care apare cel mai des in setul de date. Skewness a alcoolului este de 0.49, ceea ce inseamna ca distributia valorilor alcoolului nu este simetrica si poate fi inclinata intr-o anumita directie. Kurtosis a alcoolului este de -0.70, ceea ce inseamna ca distributia valorilor alcoolului poate fi mai aplatizata decat o distributie normala.

Quality: Valoarea mediana a calitatii din setul de date este de 6, ceea ce inseamna ca jumatate din valorile calitatii sunt mai mari si jumatate sunt mai mici decat 6. Valoarea modala a calitatii este de 6, aceasta fiind valoarea care apare cel mai des in setul de date. Skewness a calitatii este de 0.16, ceea ce determina o distributie aplatizata.

**Analiza multivariata**

In continuare, vom realiza o matrice de corelatie pentru a vedea in ce masura anumite variabile independente sunt corelate cu cea dependenta (Quality). Pe langa acest aspect, este important de vazut si cum anumite variabile independente sunt corelate cu alte variabile de aceeasi natura. Pentru a avea un aspect mai inteligibil, am denumit variabilele (aciditatea fixă, aciditatea volatilă, aciditatea citrică, zahărul rezidual, clorura, dioxidul de sulf liber, dioxidul de sulf total, densitatea, pH-ul, sulfații, alcoolul și calitatea) cu I1, I2, I3, … I12.

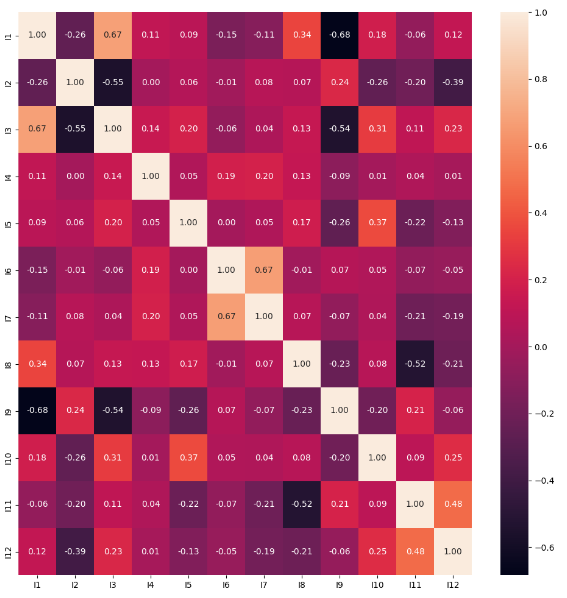
Rezultatul obtinut pentru datele legate de vin rosu:

Figura 1. Matrice de corelatie pentru vin rosu

Se observa ca I11 si I12 au un grad de corelatie destul de mare si direct (0.48), si I2 si I12 au un grad de corelatie destul de mare si invers (-0.39). Aceste informatii privesc relatia dintre variabilele independente si cea dependenta.

Apoi, putem observa ca intre I1 si I3 este un grad de corelatie foarte ridicat si direct (0.67) si intre I1 si I9, de asemeneni, un grad de corelatie ridicat dar invers (-0.68).

In plus, I2 si I3 sunt invers corelate (-0.55). la fel si I3 cu I9 (-0.54).

Pe langa aceste aspecte, mai exista o corelatie destul de mare si inversa intre I8 si I11 (-0.52).

Din moment ce nu exista valori mai mari de 0.8, sau mai mici de -0.8, nu exista riscul unui caz de multicoloniaritate care sa afecteze negativ rezultatele analizei de regresie.

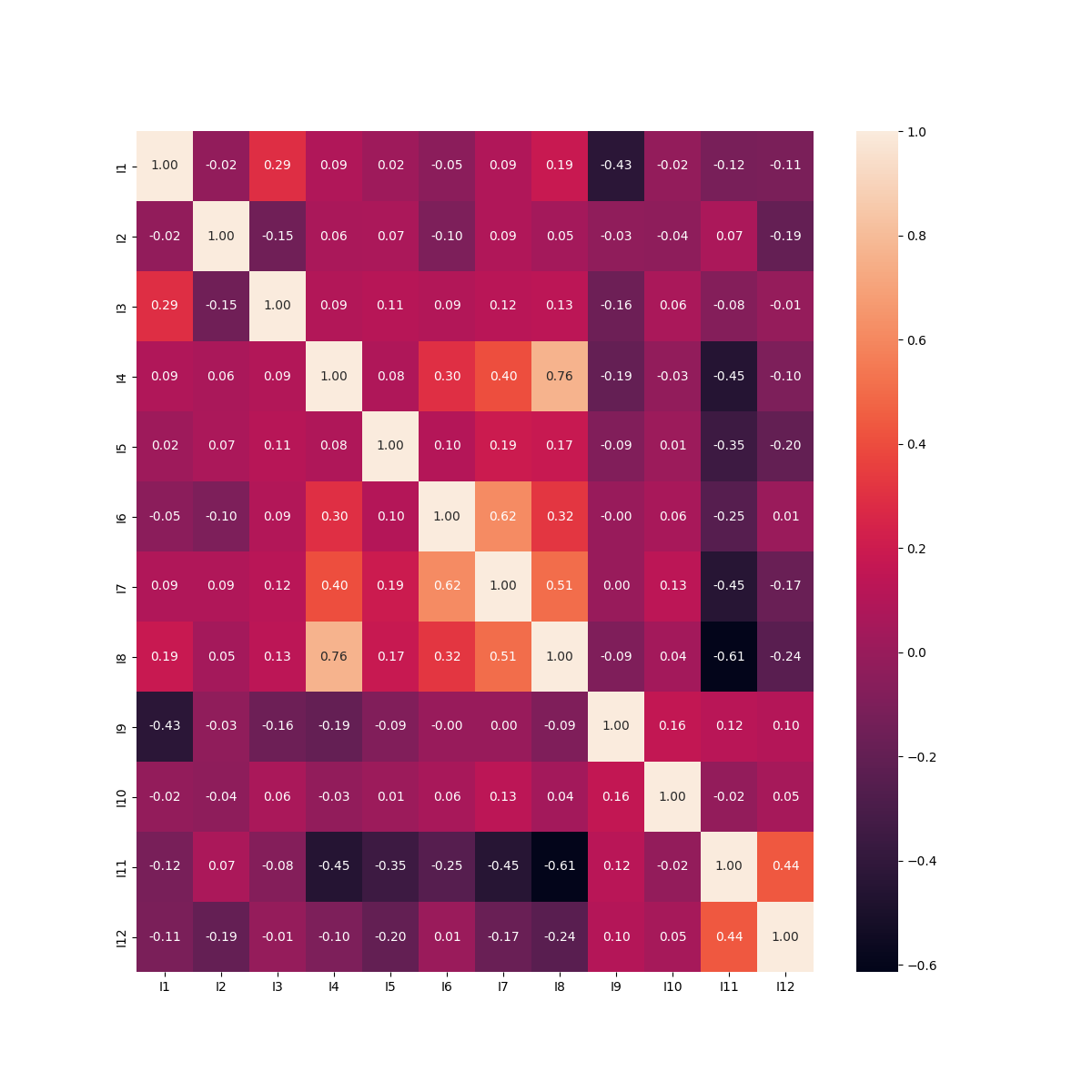
Rezultatul obtinut pentru datele legate de vin alb:

Figura 2. Matrice de corelatie pentru vin alb

Privind datele pentru vin alb, variabila dependenta, I12, cunoaste un grad destul de ridicat de corelatie (0.44) cu I11. In rest, se regasesc valori mari de corelatie: intre I1 si I9 (-0.43), intre I4 si I11 (-0.45), intre I4 si I8 (0.76), intre I6 si I7 (0.62), intre I7 si I8 (0.51), intre I7 si I11 (-0.45), intre I8 si I11 (-0.61). Avand in vedere ca nu avem valori mai mari de 0.8, nu se pune problema de multicoliniaritate care sa intervina in sens negativ in rezultatele unei analize de regresie.

**Model de regresie**

Pentru a determina in ce masura caracteristicile vinului determina calitatea acestuia, am construit un model de regresie care sa contina toate caracteristicile masurate ca variabile independente si calitatea ca variabila independenta. Rezultatele pentru aceasta analiza sunt:

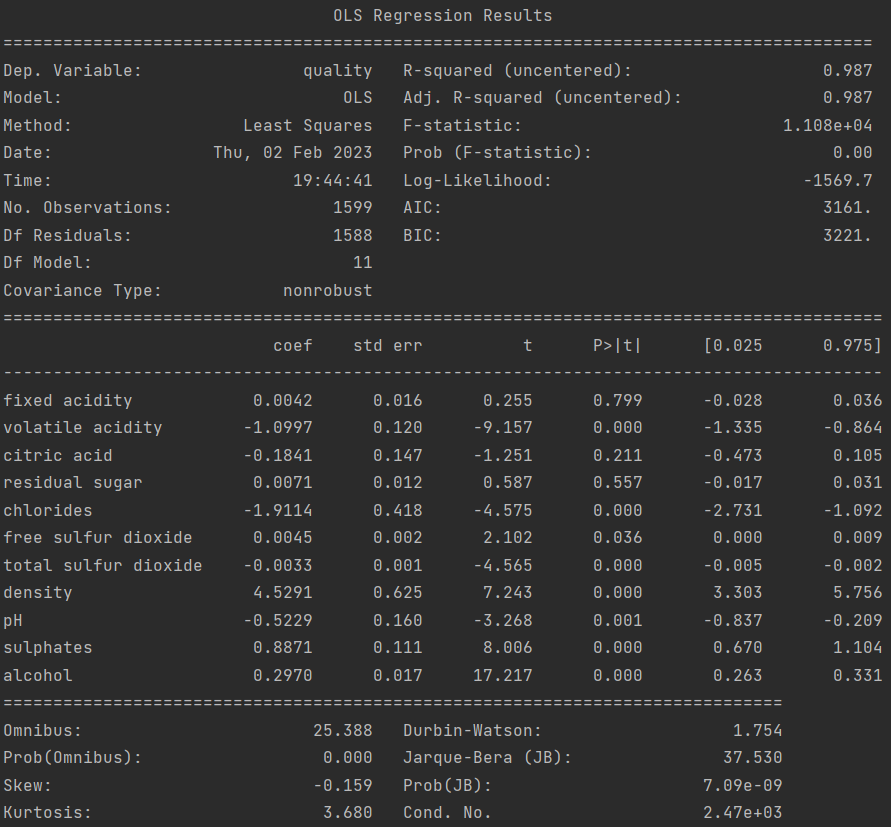


Figura 3. Rezultate initiale pentru analiza de regresie (vin rosu)

Pentru a ajusta modelul regresiei, vom lua în considerare următoarele:

Valoarea R-pătrat: Valoarea R-pătrat de 0,361 indică faptul că variabilele independente explică aproximativ 36,1% din variația variabilei dependente (calitate).

p-valorile: p-valorile variabilelor independente pot fi utilizate pentru a determina semnificația fiecărei variabile în prezicerea variabilei dependente. De exemplu, p-valoarea pentru aciditatea volatilă este aproape de 0, ceea ce sugerează că este un predictor important al calității. Pe de altă parte, p-valoarea pentru acidul citric este mai mare decât 0,05, ceea ce sugerează că poate să nu fie un predictor semnificativ al calității.

Astfel, vom elimina variabilele independente care au valoarea p mai mare decat 0.05, prin urmatoarea structura de cod:

p\_values = model.pvalues  
significant\_variables = [col for col, p\_val in zip(X.columns, p\_values) if p\_val < 0.05]  
X = df[significant\_variables]  
model = sm.OLS(y, X).fit()  
print(model.summary())

Rezultatele pentru analiza de regresie ulterioara sunt afisate in figura 4, unde putem observa ca, din nou, apare o valoare a carei p-value este mai mare decat 0.05, si anume “free sulfur oxide”, ceea ce inseamna ca analiza mai trebuie remodelata inca o data, eliminand acea variabila independenta. De altfel, se poate observa ca valoarea pentru R-squared este de 0.986, ceea ce inseamna ca variabilele independente explica 98.6% din valoarea independenta.

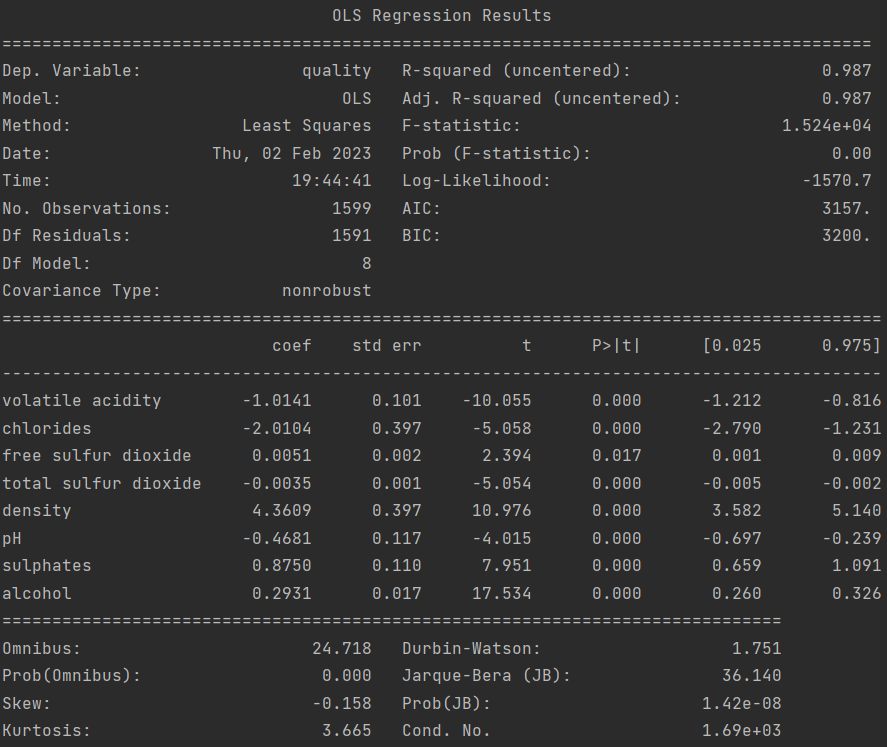
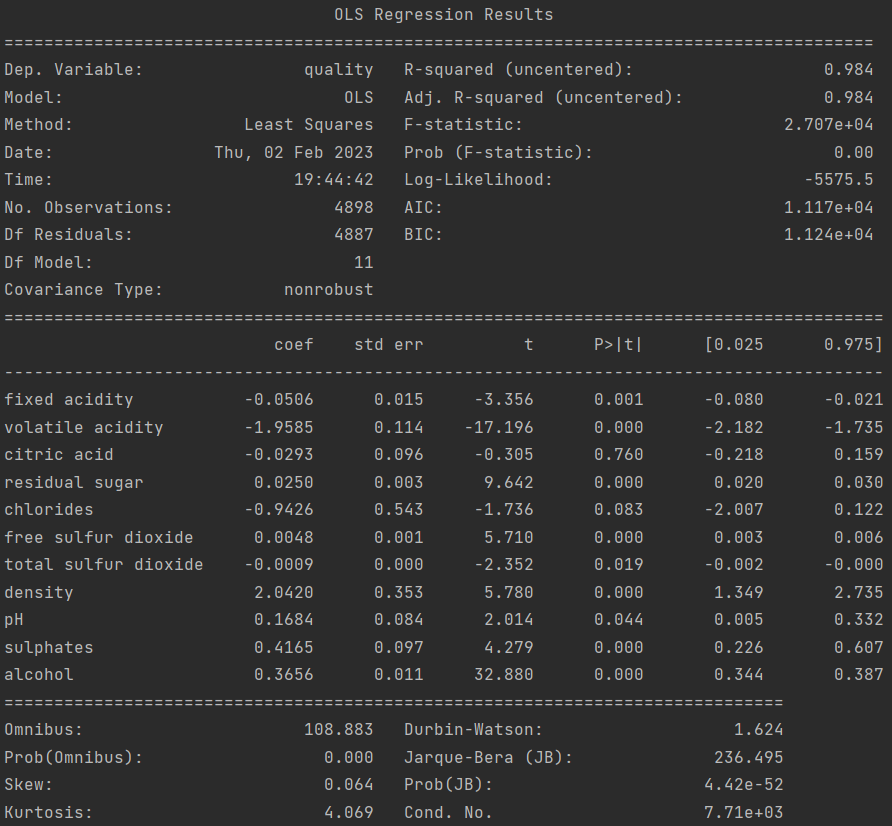


Figura 4. Rezultate pentru analiza de regresie ajustata (vin rosu)

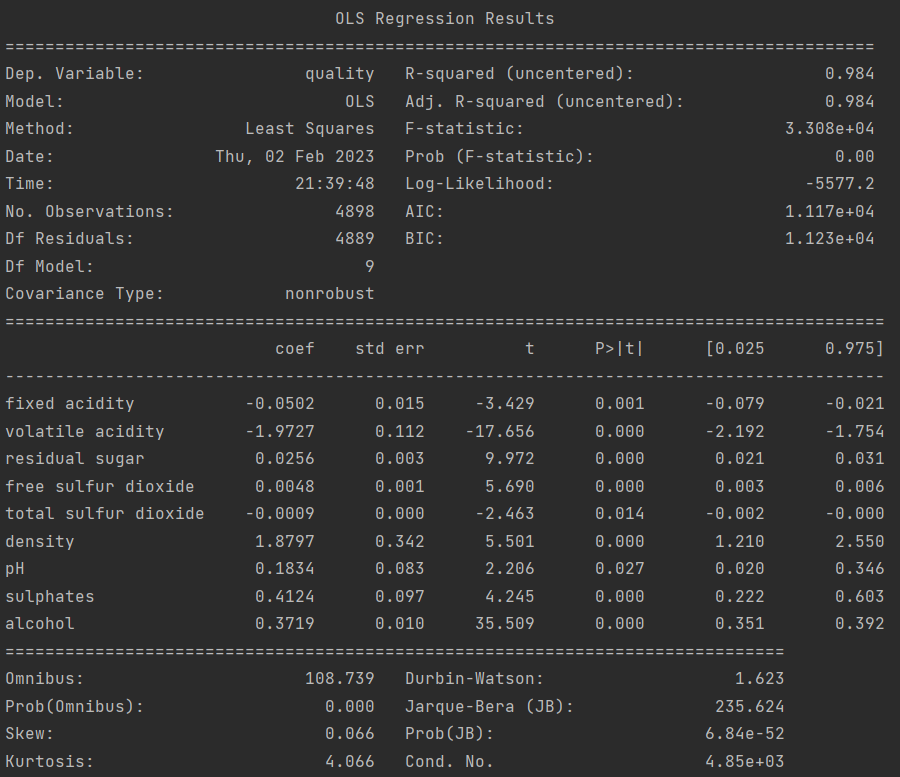
**Ecuatia de regresie rezultata este:**

**quality = -1.0141 \* volatile acidity - 2.0104 \* chlorides + 0.0051 \* free sulfur dioxide - 0.0035 \* total sulfur dioxide + 4.3609 \* density - 0.4681 \* pH + 0.8750 \* sulphates + 0.2931 \* alcohol**

In continuare, vom urma aceiasi pasi pentru a obtine ecuatia de regresie pentru calitatea vinului alb.

 Figura 5. Rezultatele analizei de regresie (vin alb)

Avand in vedere existenta valorilor “p-values” mai mare decat 0.05, vom rula aceeasi structura de cod si pentru acest set de date, astfel incat sa ramana doar variabilele independente care au o importanta semnificativa.

 Figura 6. Rezultatele analizei de regresie ajustata (vin alb)

Ecuatia de regresie rezultata este: quality = -0.0502 \* fixed acidity - 1.9727 \* volatile acidity + 0.0256 \* residual sugar + 0.0048 \* free sulfur dioxide - 0.0009 \* total sulfur dioxide + 1.8797 \* density + 0.1834 \* pH + 0.4124 \* sulphates + 0.3719 \* alcohol

**CONCLUZII**

Ambele modele de regresie prezinta o predictie a calitatii vinului in functie de diferite caracteristici ale acestuia, cum ar fi aciditatea, clorura, dioxidul de sulf, densitatea, pH-ul, sulfatii si continutul de alcool.

Pentru vinul rosu, modelele arata ca aciditatea volatila, clorura si dioxidul de sulf total au un impact negativ asupra calitatii vinului, in timp ce densitatea, sulfatii si continutul de alcool au un impact pozitiv asupra calitatii.

Pentru vinul alb, modelele arata ca aciditatea fixa si aciditatea volatila au un impact negativ asupra calitatii, in timp ce zaharul rezidual, densitatea, pH-ul, sulfatii si continutul de alcool au un impact pozitiv asupra calitatii.

In concluzie, aceste modele de regresie sugereaza ca exista o serie de factori care afecteaza calitatea vinului, fie pozitiv sau negativ, si ca aceste factori pot fi diferiti pentru vinul rosu si pentru vinul alb. Este important sa se ia in considerare aceste modele atunci cand se evalueaza calitatea vinurilor sau atunci cand se proiecteaza noi tipuri de vinuri, indicand cum trebuie sa fie calculate caracteristicile acestuia pentru a rezulta un vin de calitate.