****

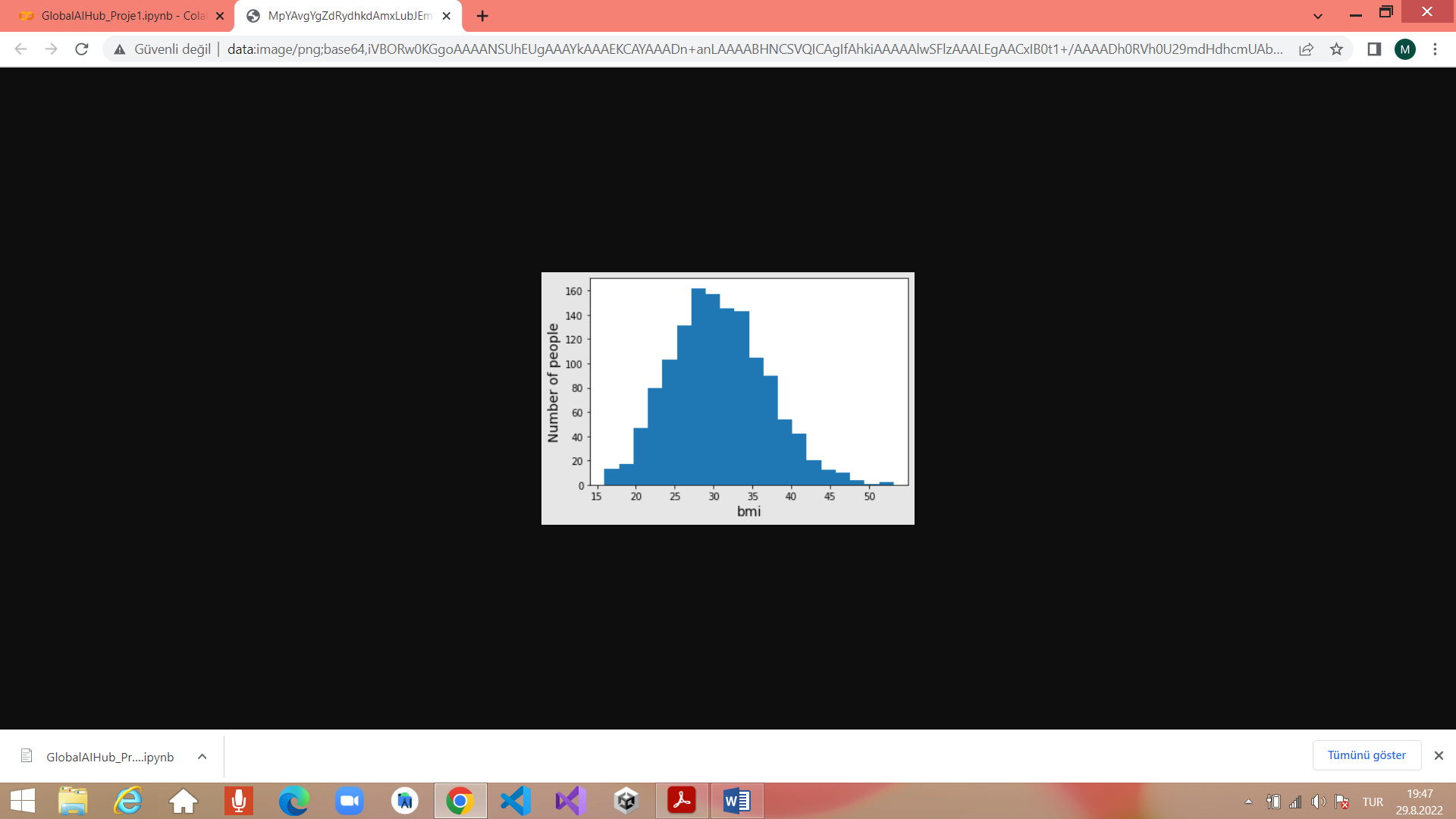
**Medikal Maliyet Analizi**

Bu projede, verilen değişkenlere göre bir kişinin sağlık sigortasının yaklaşık ne kadar masraflı olacağı tahmin edilmeye çalışılmıştır. Öncelikle Pandas, numpy, Seaborn, Matplotlib ve Sklearn kütüphaneleri eklenmiştir. Daha sonra veri seti indirilmiş ve incelenmeye başlanmıştır. Veri setinde eksik bir veri olmadığını gözlemlenmiştir.

**Keşifsel Veri Analizi Yapılması**

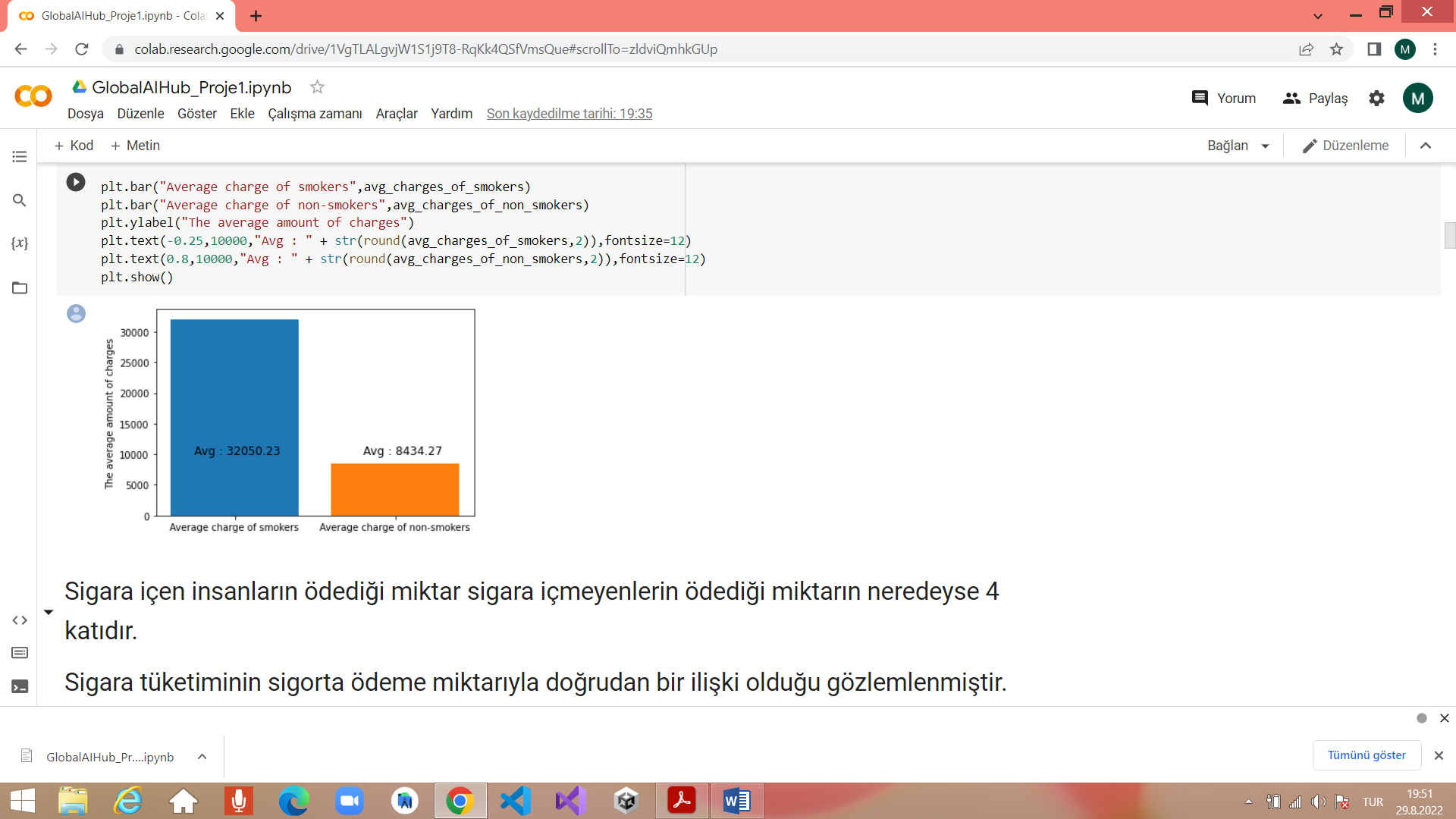
1. Bmi(Vücut Kitle İndeksi)’nin dağılımını incelenmesi

Vücut kitle indeksi ortalaması 30.66, standard sapma değeri 6.1 bulunmuştur. Bmi dağılımı Gauss Dağılımı(Normal Dağılım) şeklindedir. Vücut kitle indeksinde obezite yoğunlaşması görülmüştür.



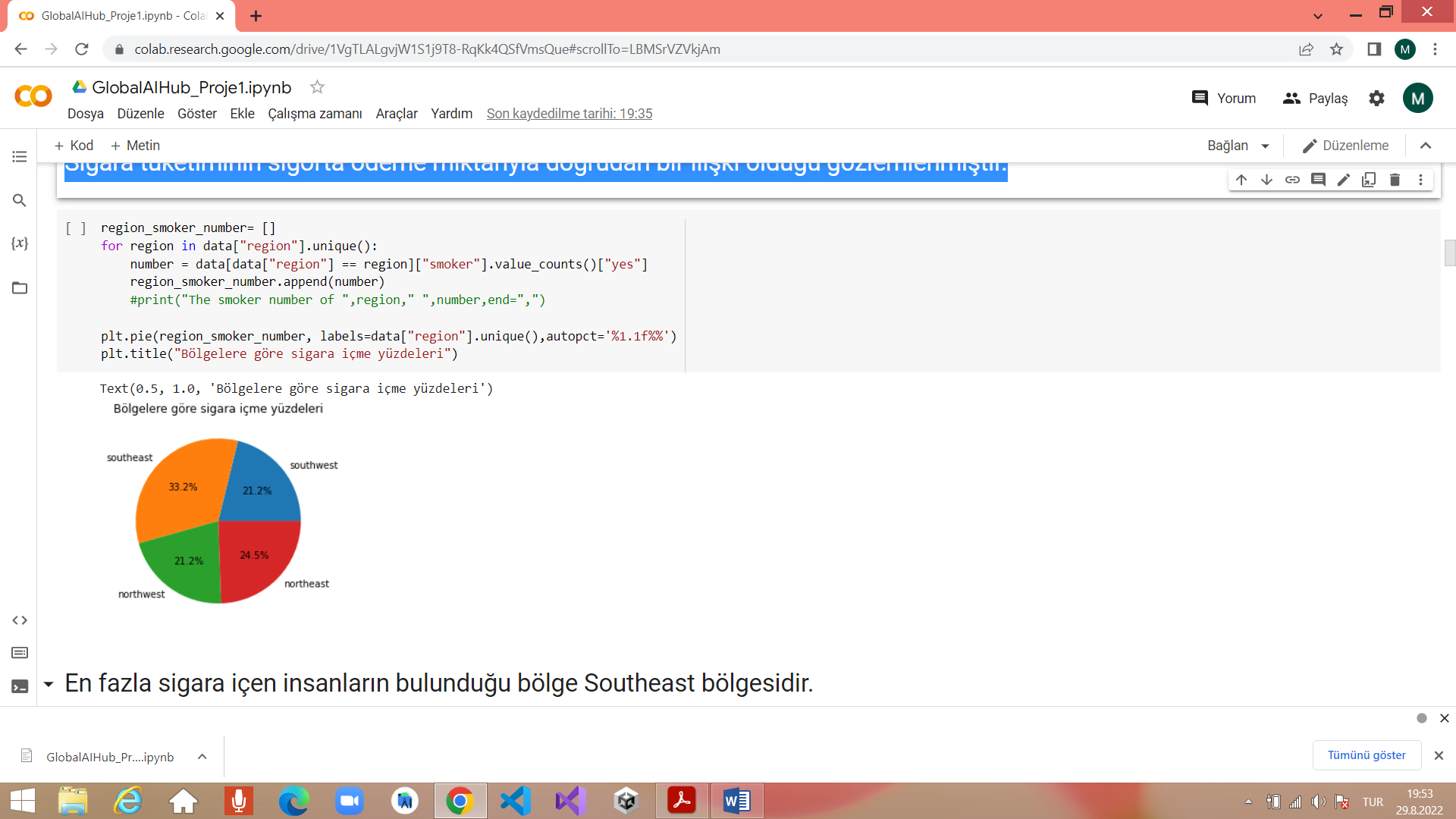
2. “smoker” ile “charges” arasındaki ilişkinin incelenmesi

Sigara içen insanların ödediği miktar sigara içmeyenlerin ödediği miktarın neredeyse 4 katıdır. Sigara tüketiminin sigorta ödeme miktarıyla doğrudan bir ilişki olduğu gözlemlenmiştir.



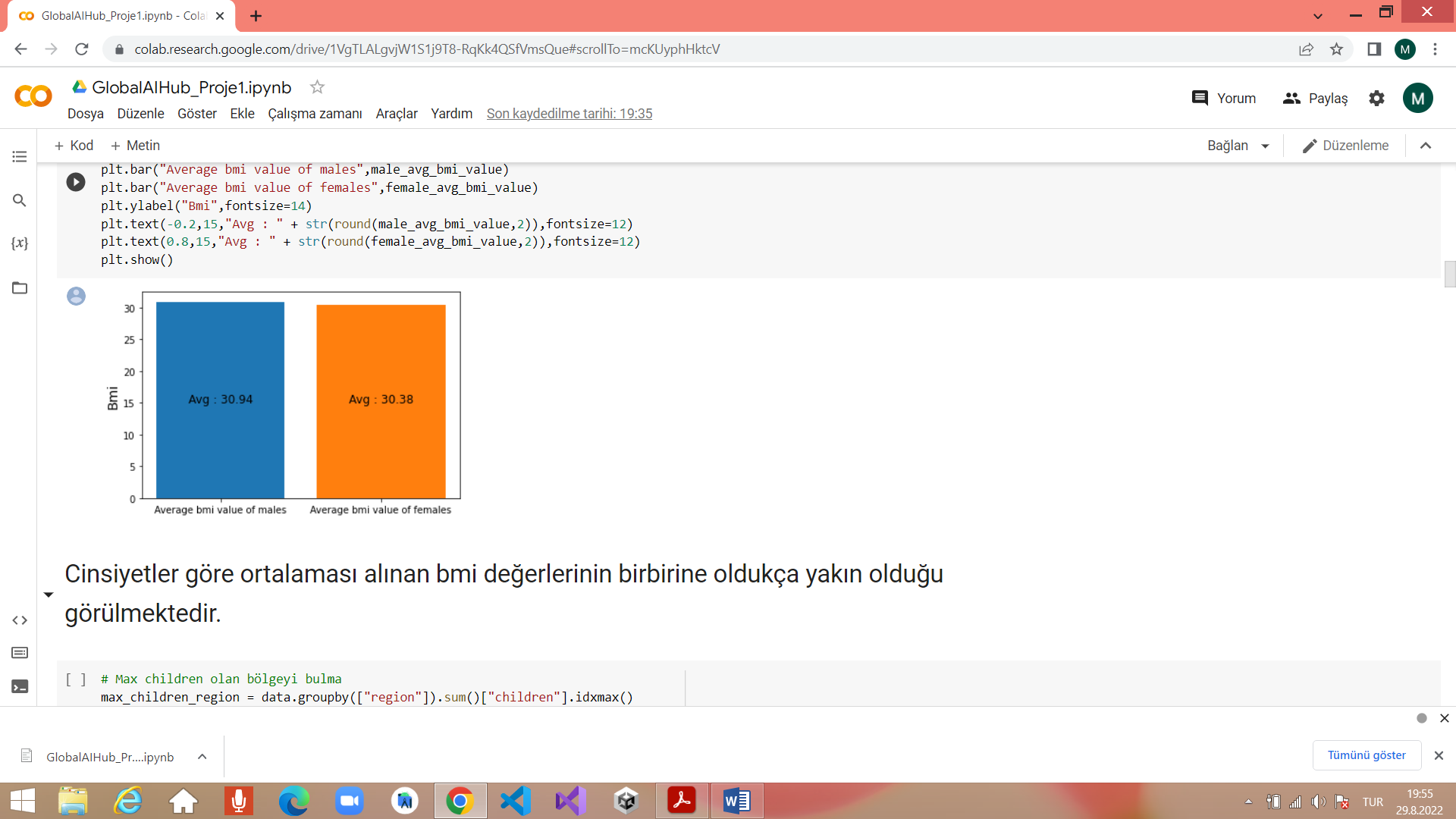
3. “smoker” (Sigara tüketen) ile “region”(Bölge) arasındaki ilişkinin incelenmesi.

En fazla sigara içen insanların bulunduğu bölge Southeast bölgesidir.



4. “bmi” ile “sex”(Cinsiyet) arasındaki ilişkinin incelenmesi.

Cinsiyetler göre ortalaması alınan bmi değerlerinin birbirine oldukça yakın olduğu görülmektedir.

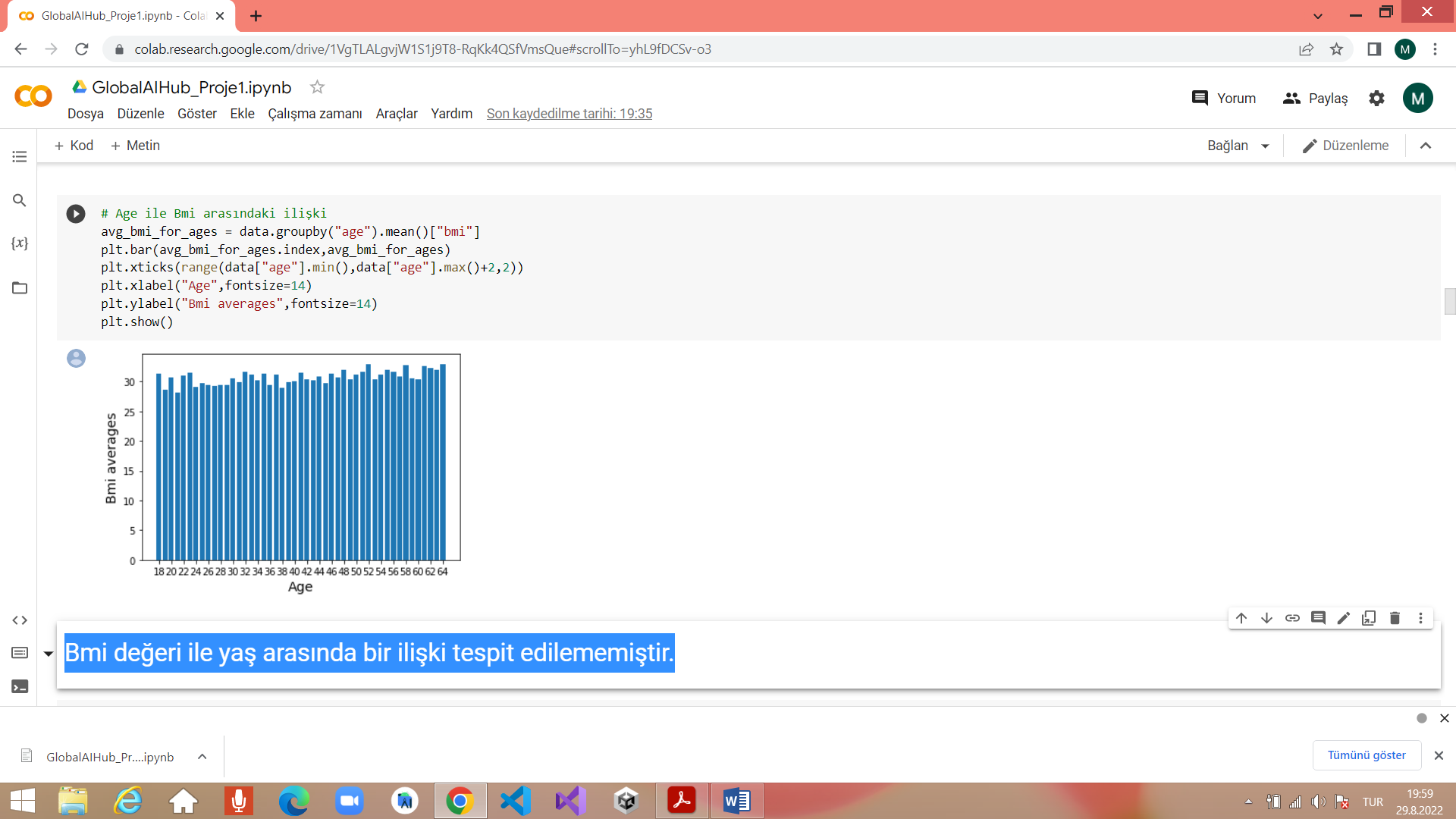
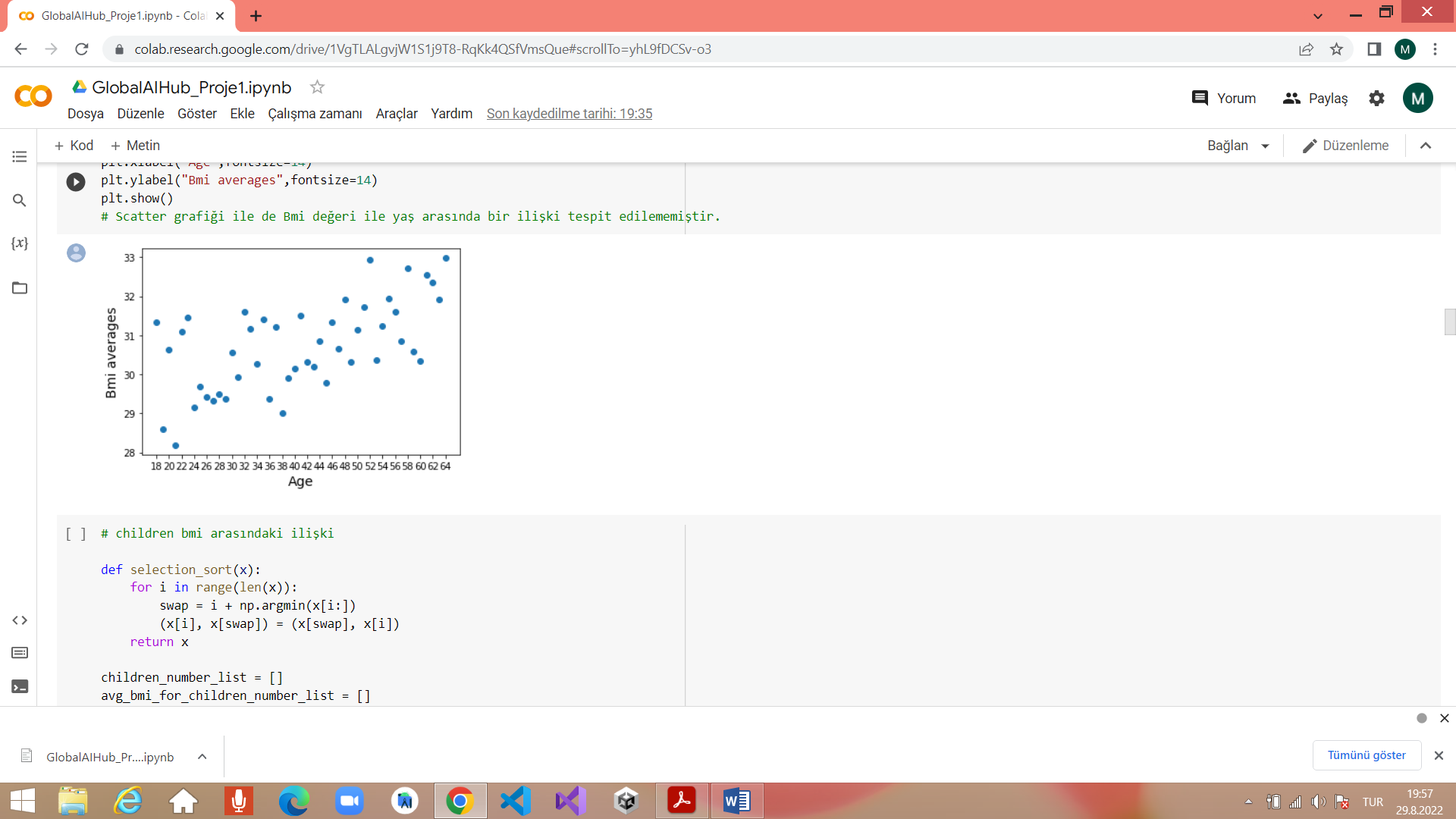


5. En çok “children”’a sahip “region” ın bulunması.

En çok çocuk sayısının olduğu bölge: southeast

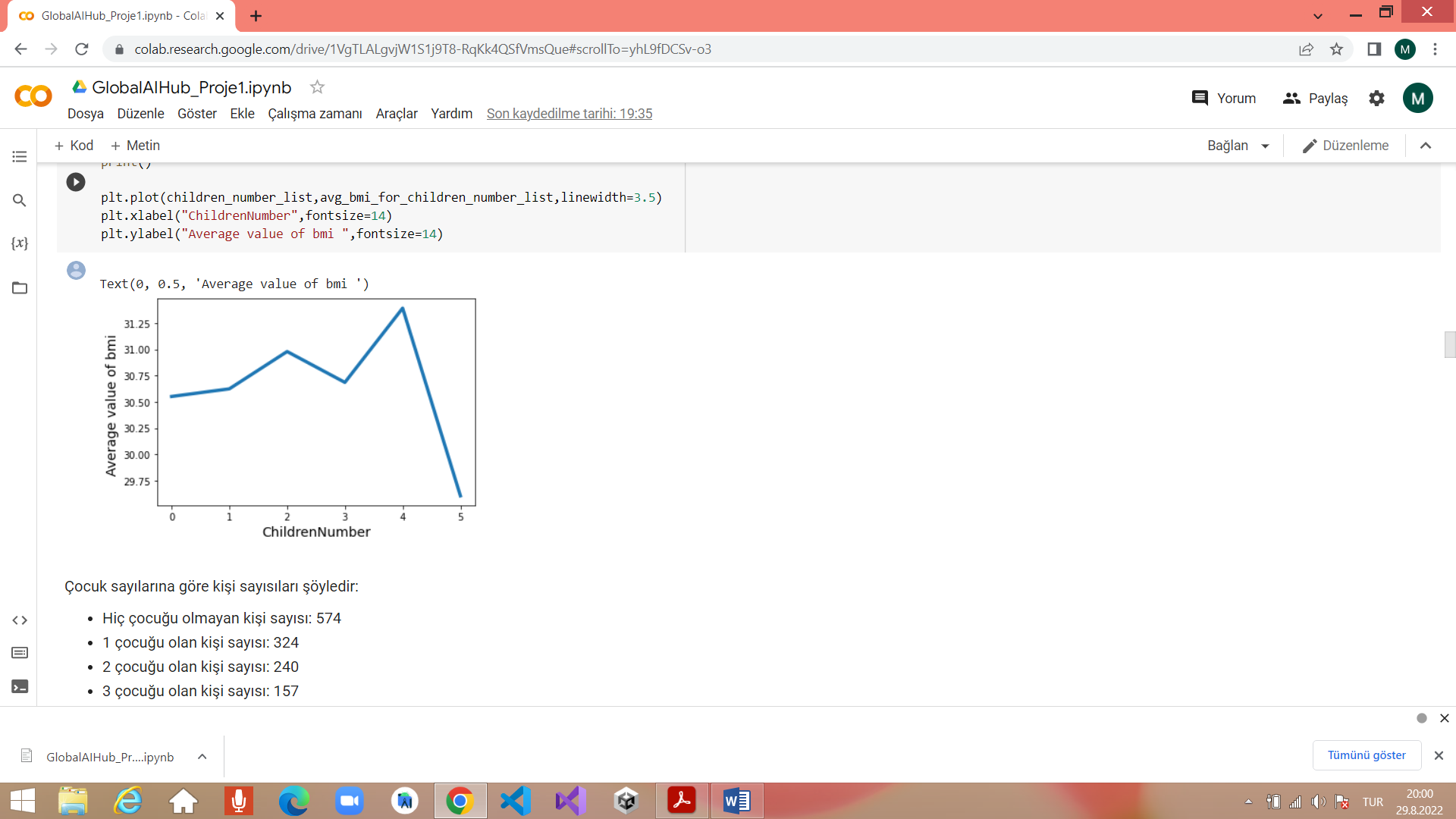
6. “Age” ile “bmi” arasındaki ilişkinin incelenmesi.

Bmi değeri ile yaş arasında bir ilişki tespit edilememiştir.



7. “bmi” ile “children” arasındaki ilişkinin incelenmesi.

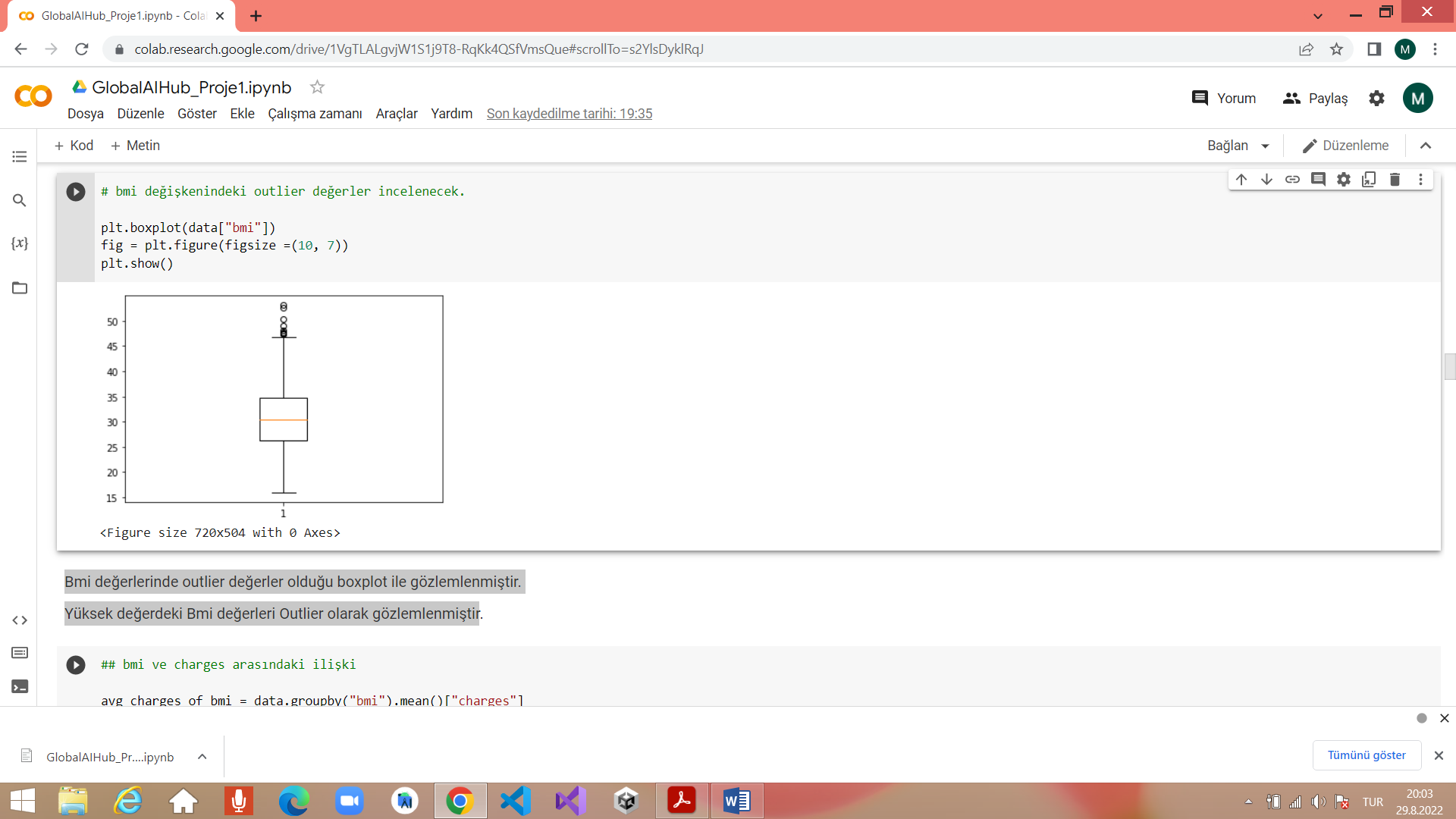
Grafiğe göre 4 ve 5 çocuğu olan kişi sayıları diğer sayıda çocuk sahibi olan kişi sayılarından çok daha az olduğu için dikkate alınmamıştır. 0, 1 ve 3 çocuğu olanların ortalama Bmi değerleri birbirine yakın iken, 2 çocuğu olan kişilerin Bmi değerleri nispeten daha yüksektir.



8. “bmi” değişkeninde outlier incelenmesi.

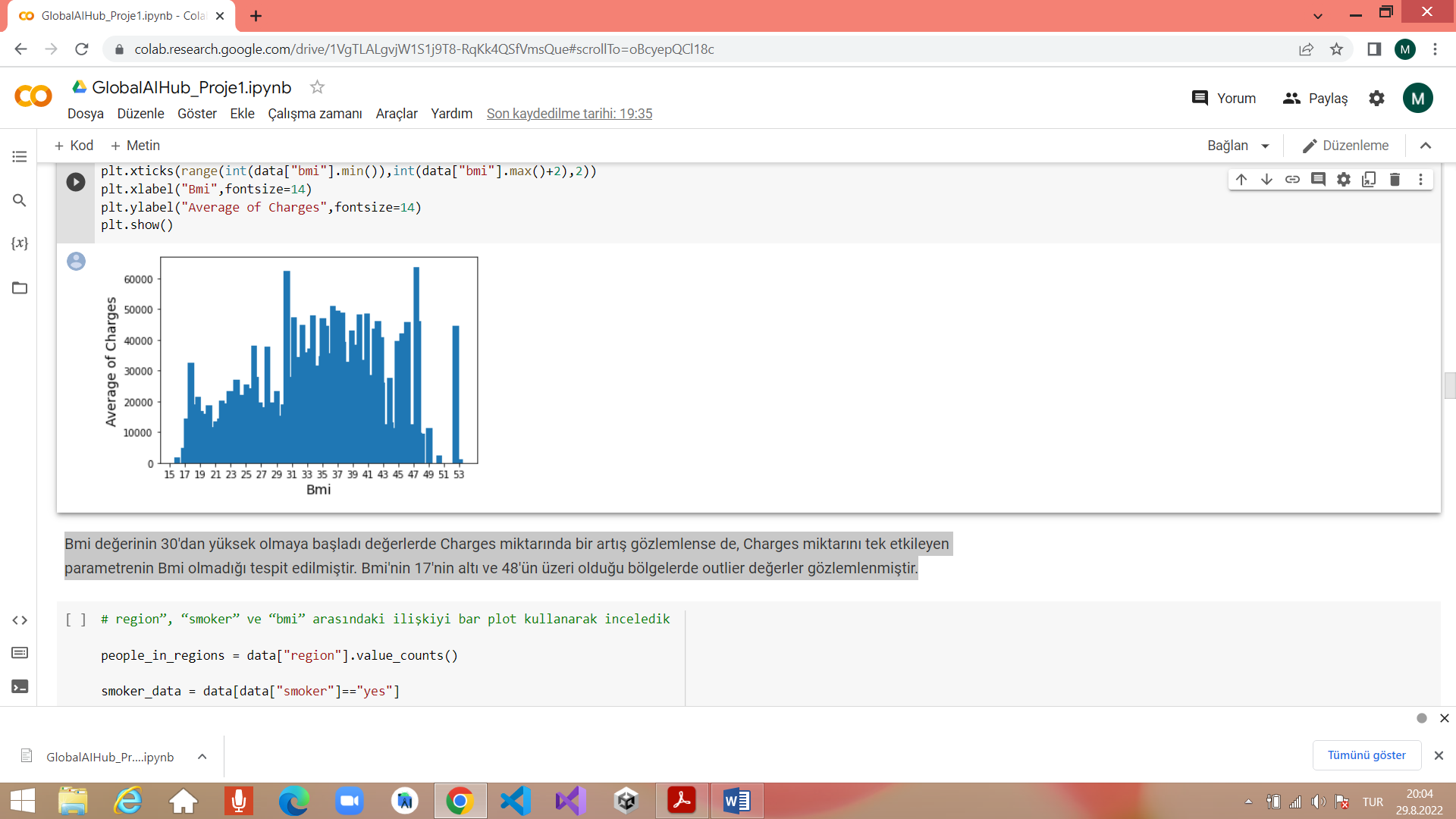
Bmi değerlerinde outlier değerler olduğu boxplot ile gözlemlenmiştir.

Yüksek değerdeki Bmi değerleri Outlier olarak gözlemlenmiştir

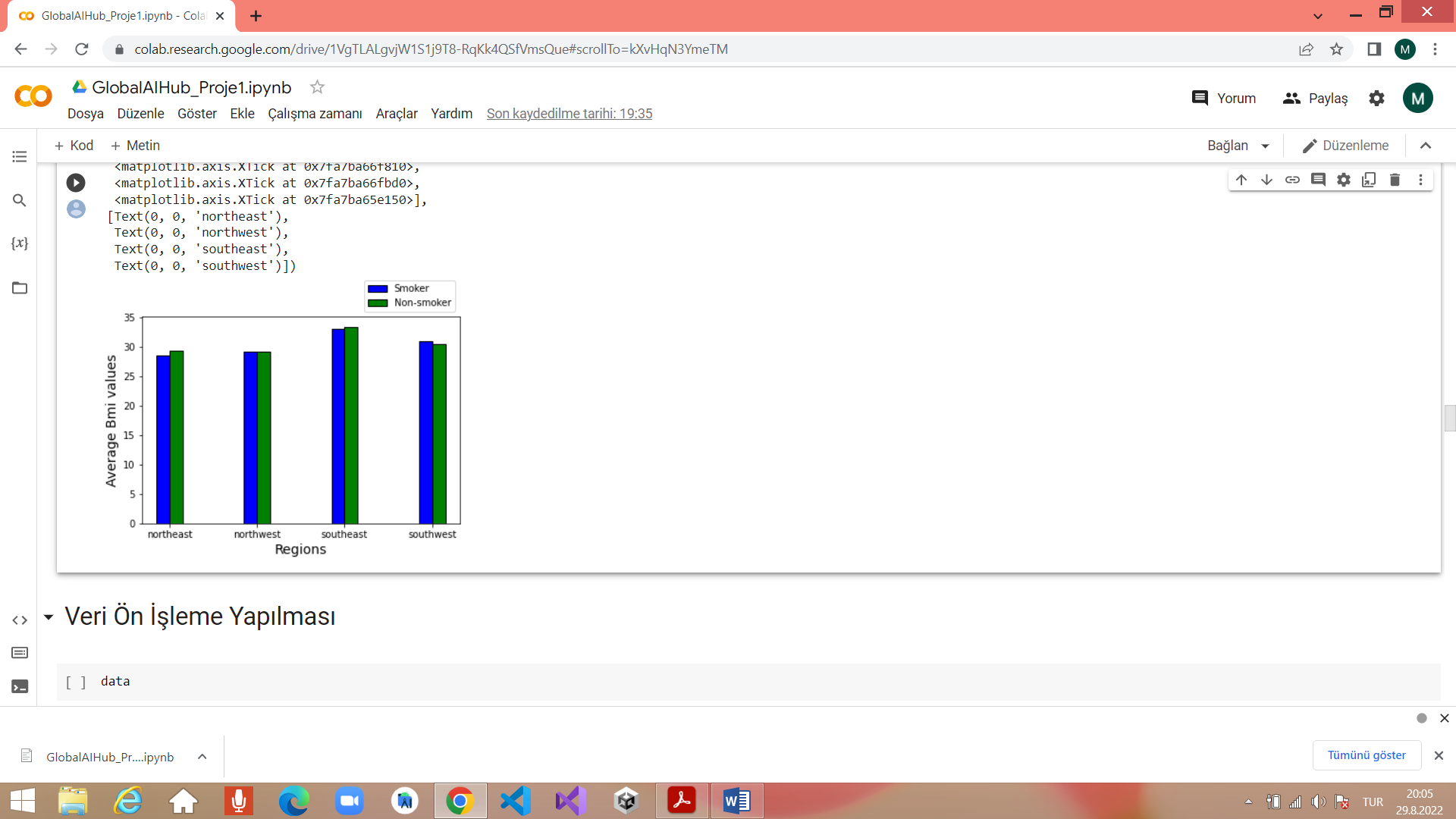


9. “bmi” ile “charges” arasındaki ilişkinin incelenmesi

Bmi değerinin 30'dan yüksek olmaya başladı değerlerde Charges miktarında bir artış gözlemlense de, Charges miktarını tek etkileyen parametrenin Bmi olmadığı tespit edilmiştir. Bmi'nin 17'nin altı ve 48'ün üzeri olduğu bölgelerde outlier değerler gözlemlenmiştir.



10. “region”, “smoker” ve “bmi” arasındaki ilişkiyi bar plot kullanarak incelenmiştir.



**Veri Ön İşleme Yapılması**

Kategorik değişkenleri düzenlemek için Label ve One-Hot Encoding teknikleri kullanılmıştır. Cinsiyet ve smoker sütunlarına Label Encoding, region sütununa One-Hot Encoding uygulanmıştır.

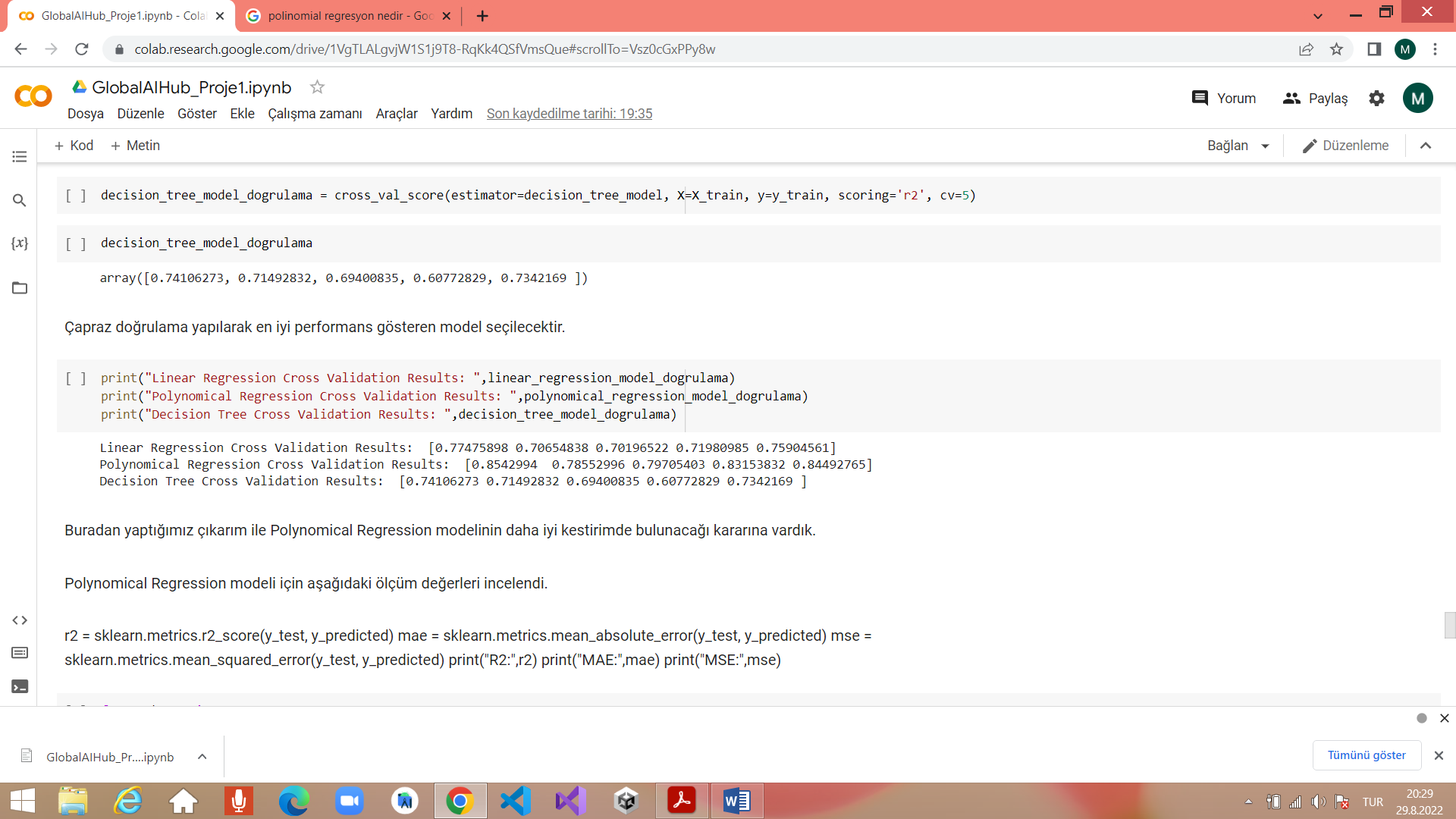
Elde edilen sonuçlara göre Bmi, Smoker olup olmama durumu, hangi bölgede yaşadığı, önemli parametrelerdir.

Bmi 17'den küçük, 46'dan büyük değerleri çıkarılmış, Children sayısı da 4 ve 5 olan toplam 43 satır silinmiştir.

Veri setinizi X\_train,X\_test, y\_train, y\_test olacak şekilde bölünmüştür. Veri setini normalize ederek ölçeklenmiştir.

**Model Seçme**

Lineer Regresyon, Polinomsal Regresyon ve Karar Ağacı modelleri uygulanmıştır. Çapraz doğrulama yapılarak en iyi performans gösteren Polinomsal Regresyon kullanılmasına karar verilmiştir.



**Hiper-parametre Optimizasyonu**

Bir önceki adımda seçilen modelin hiper-parametreleri ve Grid Search ile parametreleri optimize edilmiştir. Best score değeri 0.84 bulunmuştur.

**Modeli Değerlendirme**

Regresyon modeli değerlendirme metriklerini kullanarak optimize edilmiş olan modelin değerlendirmesi yapılmıştır. Aşağıdaki değerlere ulaşılmıştır.

