Mall Customer Segmentation EDA ve İstatistiksel Analiz Planı

Elimizdeki veriler, tamamıyla sanal bir biçimde oluşturulmuştur ve gerçek veriler değildir. Elimizdeki veriler için %95 güven aralığında, gücü 0,90 ve Cohen's d 0,5 alırsak bu durumda elimizdeki 200 kişilik örneklem, 1250 kişilik bir evreni yansıtmaktadır. Buna göre örneklemimizin rastgele ve bağımsız örnekleme metoduyla oluşturulduğu varsayılmaktadır.

Tüm bu verilere göre elimizde örneklemimizin cinsiyeti, yaşı, yıllık geliri ve harcama skorları bulunmaktadır. Elimizdeki değişkenlerin veri türlerini şöyle sıralayacağız:

- Cinsiyet → Nominal (Kategorik)
- Yaş → Ratio (Sayısal)
- Yıllık Gelir → Ratio (Sayısal)
- Harcama S. → Interval (Sayısal) (Aslında nominal/interval arası)

Verilerimizi daha iyi tanımak adına öncelikli olarak tanımlayıcı istatistiklerimizi gerçekleştireceğiz. Buna göre cinsiyet, yıllık gelir ve harcama skoru değişkenlerimizin ortalama, medyan, mod, varyans, standart sapma, basıklık, çarpıklık, minimum, maksimum, çeyreklik dağılımı (kartil) değerlerine bakacağız. Bu veriler ışığında verilerimizin normalliğini ön test olarak kontrol etmiş olacağız (elimizdeki veri bir sosyal bilim verisi, son zamanlardaki istatistiki gelişmelerde standart sapma değerinin +/-1 arasında yer alması normal dağılım için yeterli kabul ediliyor ancak biz, normallik iddiamızı güçlendirmek için diğer sonuçlara da bakacağız). Sonrasında bu değerlerin standart hatasını hesaplayıp normallik varsayımımızı kuvvetlendirecek ve sonrasında grafiklere bakacağız. Histogramlarına bakarak normalliğine dair varsayımımızı kuvvetlendirmeye çalışacağız. Sonrasında kutu-bıyık grafiğine bakarak uç değerleri tespit etmeye çalışacağız. Uç değerleri, regresyon testleri için veriden atmayı ya da manipüle etmeyi düşüneceğiz. Sonrasında hipotezlerimize geçeceğiz.

Daha sonrasında regresyon ve varyans analizi yapabilmek adına yaş, yıllık gelir ve harcama skoru değişkenlerimizi nominal veriye dönüştürmeye çalışacağız. Nominal veri türüne dönüştürürken mevcut sayısal veriyle oluşturduğumuz kategorik veri için aralarında regresyon testi (sonrasında en küçük kareler yöntemine başvurarak) ve AIC/BIC skorlarına bakarak kategorilerimizin sayısal veriyi doğru temsil edip etmediğini kontrol edecek ve ona göre kategorik verimizi oluşturacağız. Ayrıca regresyon testi için korelasyon analizleri de uygulayacağız (AIC/BIC>6 olması beklenecek).

Sonucunda hipotezlerimiz için şu kategoriler oluşacak:

- Cinsiyet → Nominal (Kategorik)
- Yaş → Ratio (Sayısal)
- Yıllık Gelir → Ratio (Sayısal)
- Harcama S. → Interval (Sayısal) (Aslında nominal/interval arası)
- Grup. Yaş → Nominal (Kategorik)
- Grup. Gel. → Nominal (Kategorik)

Daha sonrasında elimizdeki verilerle Kruskal-Wallis, Shapiro-Wilk normallik testi ve Levene homojenlik testi yapılarak hipotezlerimiz için sonuç üretilecek.

Hipotezler:

Aşağıda alternatif hipotezlerimizi (h1) ve onları gerçekleştirmek için gerekli istatistiki testleri liste halinde paylaşacak ve normallik varsayımının sağlanamadığı durumlarda alternatifleri en altta parantez içerisinde belirtilecektir.

1. Örneklemimizin harcama davranışının evren için ortalama/varyans güven aralığı oluşturma:
   T testi uygulanacak ve elde edilen değer, nokta tahminiyle kıyaslanarak bir değer oluşturularak evrenin tanımlayıcı değerlerini oluşturmak.
2. Yaş arttıkça harcama skoru ve yıllık gelir artar:
   Yaş, harcamanın ve gelirin nedeni varsayımı söz konusudur. Çoklu regresyon yapılacak ve sonrasında nedenselliği kanıtlanmak için SEM testi yapılacak. Oluşturduğumuz kategorik değişkenlerin doğruluğunu ispatlamak için ayrıca kategorik değişkenlerle lojistik regresyon da yapılacak ve iki regresyonun karesi karşılaştırılacak.
3. Cinsiyete göre harcama skoru ve yıllık gelir değişir.
   Tek Yönlü MANOVA testi yapılarak hipotezler oluşturulacak. Sonuca göre post-hoc testleri (Bonferroni, Tukey tercih edilecek) yapılarak anlamlılığı bozan parçalar tespit edilecek ve sonrasında gerekirse ANOVA testi ve ANCOVA testi gerçekleştirilerek hipotezin doğruluğu tam tespit edilecek. MANOVA sonucunda Wilks' Lambda, Pillai's Trace ve Hotelling's Trace değerleri de incelenecek.
   (Normallik sağlanamadığı durumda non-parametrik alternatifi, oluşturulan evren varyansı yeteri kadar güvenilir olmadığında Welch alternatifleri uygulanacaktır.)
4. Yaş grupları arasında anlamlı bir gelir ve harcama farkı vardır:
   Tek yönlü MANOVA testi yapılacak. Daha sonrasında oluşturulan alt hipotezlere göre ANOVA ve ANCOVA testleri uygulanacak. Eğer doğrudan h0 reddedilemezse önce gelirin kovaryant olması, sonra da harcama farkının kovaryant olması denenerek (ANCOVA) hipotezin diğer alternatifleri test edilecek.
   Sonuca göre post-hoc testleri yapılacak ve anlamlılık yorumlanacak. MANOVA sonucunda Wilks' Lambda, Pillai's Trace ve Hotelling's Trace değerleri de incelenecek.Normallik, homojenlik, bağımsızlık sağlanamazsa non-parametrik alternatif; evren varyansı sağlanamazsa Welch alternatif kullanılacak.
5. Cinsiyete göre harcama ve gelirde farklılık vardır:
   Tek yönlü MANOVA yapılacak. Oluşturulan alt hipotezlere göre ANOVA testlerine geçilerek anlamlılık tespit edilmeye çalışılacak. Ayrıca anlamlılığın olmadığı bağımsız değişken için ANCOVA denenecek. MANOVA sonucunda Wilks' Lambda, Pillai's Trace ve Hotelling's Trace değerleri de incelenecek. ANOVA şartları sağlanamazsa non-parametrik alternatif, varyans güvenilir gelmezse Welch alternatif uygulanacak.
6. Gelir, harcamanın nedenidir:

Tek yönlü doğrusal ya da doğrusal olmayan regresyon uygulanacak.
Sonucunda h0 reddedilirse nedensel çıkarım için SEM testi yapılacak.

Hipotezlerimiz test edildikten sonra tüm sonuçlar yorumlanarak istatistiki bir sonuca varılacak ve sonuç, raporlanacak.

Üstteki aşamaların her birinin uygulanması için SQl ve R programlama dillerine başvurulacak ve daha sonrasında raporlama için Power BI kullanılacak. Power BI programında iki tür rapor bulunacak. 1. rapor, üstteki teknik bilgilere girmeden yalnızca tanımlayıcı istatistikten bazı sonuçların bulunduğu rapor olacak. 2. rapor, teknik detayların yer aldığı daha detaylı rapor olacak.

SQL tarafında JOIN ve GROUP BY yapılarıyla veri birleştirme ve özetleme işlemleri yapılacak; kategorik değişkenler CASE WHEN ve NTILE() komutlarıyla oluşturulacaktır. Aykırı değer tespiti ve manipülasyonu için de HAVING koşulları ve IQR tabanlı hesaplamalar uygulanacaktır.

R'ın ggplot2 kütüphanesiyle grafikler oluşturulacak, dplyr kütüphanesiyle SQL'den elde edilen veriler üzerinden gerekli manipülasyonlara devam edilecek, stats-car-lme4 kütüphaneleriyle ve temel komutlarıyla tanımlayıcı ve çıkarımsal analizler gerçekleştirilecektir.

R ve SQL çıktıları Power BI'a aktarılarak hem teknik hem de yönetici seviyesine uygun iki ayrı dashboard hazırlanacaktır.

Mall Customer Segmentation EDA and Statistical Analysis Plan

The dataset at hand has been completely generated artificially and does not represent real-world data. Assuming a 95% confidence interval, a power of 0.90, and a Cohen's d of 0.5, our sample of 200 individuals represents a population size of approximately 1250. Accordingly, it is assumed that the sample was created using a random and independent sampling method.

Based on this dataset, we have information on the gender, age, annual income, and spending scores of the sample. The variable types will be categorized as follows:

- Gender Nominal (Categorical)
- Age Ratio (Numerical)
- Annual Income Ratio (Numerical)
- Spending Score Interval (Numerical) (Actually between ordinal/interval)

To better understand the data, we will first perform descriptive statistical analyses. For the variables gender, annual income, and spending score, we will examine the mean, median, mode, variance, standard deviation, kurtosis, skewness, minimum, maximum, and quartile distributions. Based on these results, we will preliminarily evaluate the normality of the data (as this dataset imitates a social science dataset, recent statistical developments accept a standard deviation between +/-1 as sufficient evidence for normality; however, we will still examine other indicators to strengthen our assumption). Then, we will calculate the standard errors of these statistics to further reinforce the normality assumption and proceed to graphical analyses. We will look at histograms to support the normality assumption and boxplots to detect outliers. Outliers will be reviewed for potential removal or manipulation in preparation for regression analyses. After this stage, we will proceed to hypothesis testing.

Next, to perform regression and variance analyses, we will attempt to convert the age, annual income, and spending score variables into nominal variables. While converting numerical variables to categorical ones, we will run regression tests (followed by the least squares method) and evaluate AIC/BIC scores to determine whether the categorical variables correctly represent the numerical values, and accordingly, we will finalize our categorical variable structures. Correlation analyses will also be performed for regression testing (AIC/BIC > 6 will be expected).

As a result, the following categories will be formed for hypothesis testing:

- Gender Nominal (Categorical)
- Age Ratio (Numerical)
- Annual Income Ratio (Numerical)
- Spending Score Interval (Numerical) (Actually between ordinal/interval)
- Age Group Nominal (Categorical)
- Income Group Nominal (Categorical)

Afterward, we will apply Kruskal-Wallis, Shapiro-Wilk normality test, and Levene's homogeneity test to generate results for our hypotheses.

Hypotheses:

Below, we list our alternative hypotheses (h1) and the statistical tests that will be used to evaluate them. When normality assumptions are not met, alternative non-parametric methods will be noted in parentheses.

To estimate population mean/variance confidence intervals regarding the spending behavior of the sample:
A t-test will be performed, and the obtained values will be compared with the point estimate to generate descriptive estimates for the population.

As age increases, spending score and annual income increase:
Age is assumed to be the cause of spending and income. Multiple regression will be performed, and a SEM test will be used afterward to validate causality.

To validate the correctness of the categorical variables we created, logistic regression will also be applied, and the squares of the models will be compared.

Spending score and annual income differ by gender:
A one-way MANOVA test will be conducted. Based on the results, post-hoc tests (Bonferroni, Tukey) will be performed to determine which components cause significance, and if necessary, ANOVA and ANCOVA will be run to fully validate the hypothesis. Wilks' Lambda, Pillai's Trace, and Hotelling's Trace values will also be examined in the MANOVA results. (If normality is not met, non-parametric alternatives will be used; if population variance is not reliable, Welch alternatives will be applied.)

There are significant income and spending differences across age groups:
A one-way MANOVA test will be applied. Then, depending on the sub-hypotheses, ANOVA and ANCOVA analyses will be conducted. If H0 cannot be rejected directly, income will first be tested as a covariate, and then spending will also be tested as a covariate (ANCOVA). Post-hoc tests will then be carried out, and significance will be interpreted. Wilks' Lambda, Pillai's Trace, and Hotelling's Trace values will also be reviewed in the MANOVA results. If normality, homogeneity, or independence are not met, non-parametric alternatives will be used; if population variance is unreliable, Welch alternatives will be applied.

Spending and income differ by gender:
A one-way MANOVA will be conducted. Depending on the resulting sub-hypotheses, ANOVA tests will be performed to determine significance. If significance is not found for an independent variable, ANCOVA will be applied. Wilks' Lambda, Pillai's Trace, and Hotelling's Trace values will also be examined in the MANOVA results.
If ANOVA assumptions are not met, non-parametric alternatives will be used; if variance is unreliable, Welch alternatives will be applied.

Income is the cause of spending:
A one-way linear or nonlinear regression will be applied. If H0 is rejected, SEM testing will be used for causal inference.

After testing the hypotheses, all results will be interpreted, and a statistical conclusion will be drawn and reported.

For the implementation of all steps described above, SQL and R programming languages will be used, followed by reporting in Power BI. Power BI will include two types of reports:

1. A non-technical report that presents selected descriptive statistics without entering into technical details.
2. A more detailed technical report that includes all analytical and statistical procedures.

On the SQL side, JOIN and GROUP BY structures will be used for data merging and summarization; categorical variables will be generated using CASE WHEN and NTILE() commands. HAVING conditions and IQR-based calculations will be applied for outlier detection and manipulation.

In R, the ggplot2 library will be used to create visualizations; the dplyr library will be used for further manipulation of SQL-derived data; and descriptive/inferential analyses will be performed using the stats, car, and lme4 libraries and base functions.

R and SQL outputs will be transferred to Power BI, where two separate dashboards will be prepared for both technical audiences and managerial-level users.