**Question A: Designing A/B Tests for Launching "VibeFest"**

**1. What are the hypotheses in this study?**

**Hypotheses**:

* **Hypothesis (H1)**: Ad A will generate more clicks and ticket purchases than Ad B.
* **Null Hypothesis (H0)**: There will be no significant difference in the number of clicks or ticket purchases between Ad A and Ad B.

Hypothesis 1 or H1(+)

Clicks and tickets purchases

Ad A > Ad B

**2. For each hypothesis you wrote in QA-1, also write the null hypothesis required for the A/B testing.**

* **Hypothesis**: Ad A will generate more clicks and ticket purchases than Ad B.
  + **Null Hypothesis (H0)**: There will be no significant difference between Ad A and Ad B in generating clicks or ticket purchases.

**3. For each hypothesis you wrote in QA-1, determine the independent and dependent variable in the hypothesis QA-1.**

* **Independent Variable**: The advertisement type (Ad A or Ad B).
* **Dependent Variables**:
  1. Number of clicks.
  2. Number of ticket purchases.

**4. Explain how you would run this A/B test on Instagram.**

* **Running the ads**: Using Instagram’s ad platform to set up a split test for Ad A and Ad B, targeting users in New York and neighboring areas.
* **Target Audience**: Instagram users in New York State and nearby regions, specifically those interested in music festivals and cultural events.
* **Timeframe**: The test will run for two weeks in November.
* **Measurement**: The dependent variables (clicks and ticket purchases) will be tracked through Instagram's ad analytics system, which provides data on link clicks and sales tracking.
* **Random Assignment**: Instagram’s system can randomly assign users to view either Ad A or Ad B, ensuring an unbiased distribution of the ads across the sample.

**5. Consider your answer to QA-4, what challenges may arise during the A/B test?**

**Challenges**:

1. **Sample Bias**: Certain users may already be predisposed to purchasing tickets based on factors unrelated to the ad (e.g., previous attendance at similar festivals or brand loyalty). This bias could skew results, making it harder to isolate the true impact of the ads.
2. **Ad Fatigue**: Users may see the same ad multiple times during the two-week testing period. This overexposure can lead to reduced engagement over time, making it more difficult to determine which ad is more effective in generating clicks and purchases. Ad fatigue could disproportionately affect one of the ads, introducing unintended bias.
3. **Attribution Challenges**: Tracking ticket purchases back to the exact ad that led to the conversion can be difficult. Users might see an ad on Instagram but purchase tickets through a different channel later, or after being exposed to multiple ads. Proper attribution is crucial to ensure accurate results in measuring ticket purchases.
4. **Platform Limitations**: Instagram’s algorithm may not distribute the ads equally to all user segments. Although the platform aims for random assignment, external factors like user engagement or content preferences could lead to unequal exposure across demographics or regions, potentially skewing the results.
5. **External Factors**: Seasonality, competing events, or general interest in festivals at the time of testing could affect engagement. For example, if there is a competing festival with heavy marketing during the test, users may be less inclined to engage with either ad. These factors could interfere with the effectiveness of the ads and the results of the test, making it difficult to determine the ad’s true impact on ticket purchases.

**6. For each hypothesis you wrote in QA-1, name the statistical analysis you need to run to test the hypothesis.**

 A **Chi-square test** or a **two-proportion Z-test** would be appropriate for analyzing the differences between the two ads (Ad A and Ad B) in terms of clicks and ticket purchases. These tests help to determine whether there is a statistically significant difference between the two groups.

* **Chi-square Test**: This test is ideal for analyzing categorical data (e.g., the number of users who clicked or purchased tickets vs. those who did not). By comparing the observed counts (clicks or purchases) from Ad A and Ad B to the expected counts, the Chi-square test can determine whether the differences in engagement between the two ads are statistically significant.
* **Two-proportion Z-test**: If we are analyzing proportions (e.g., the percentage of users who clicked or purchased tickets out of the total number of users exposed to each ad), a two-proportion Z-test may be used. This test compares the proportion of clicks or purchases from each ad to identify whether the difference between the two ads is statistically significant.

 **Why These Tests**: Both tests are designed to test hypotheses involving categorical outcomes (e.g., whether someone clicked or purchased). The goal is to determine whether any observed differences in clicks and purchases between the two ads are large enough to be unlikely due to chance, thereby providing evidence that one ad performs better than the other.

* If the **p-value** from these tests is less than a predefined significance level (usually 0.05), we can reject the null hypothesis and conclude that Ad A or Ad B performs significantly better in generating clicks or ticket purchases.

**1. What are the hypotheses in this A/B test?**

* **Hypothesis 1 (H1)**: Customers who experience a $1 price increase will have fewer cancellations compared to those who experience a $2 price increase.
* **Null Hypothesis (H0)**: There will be no significant difference in the cancellation rates between the customers experiencing a $1 increase and those experiencing a $2 increase.

Hypothesis Hypothesis 1 or H1(+)

Fewer Cancellations

$1 Price Increase vs. $2 Price Increase

**2. Null Hypothesis for Each Hypothesis**

* For the hypothesis stated above, the **Null Hypothesis (H0)** is that the proportion of cancellations for the $1 increase group (Group A) will be the same as the proportion of cancellations for the $2 increase group (Group B).

**3. Independent and Dependent Variables**

* **Independent Variable**: The price increase ($1 increase for Group A vs. $2 increase for Group B).
* **Dependent Variable**: Cancellation behavior of the customers (whether they canceled their subscription or not in April and May).

**Question 4: Before running any analysis, clean the dataset. In your Word submission briefly explain the steps you took to clean the data. In your Python file submission include the code you used to clean the dataset. (2 pts)**

**Answer**:

 **Data Loading**:

* Loaded the dataset using pandas.

 **Data Checking**:

* **Checked for Missing Values**:
  + Used df.isnull().sum() to confirm there are **no missing values** in any of the columns.
* **Verified Data Types**:
  + Ensured that the data types are appropriate for analysis:
    - ID: int64
    - Condition: object (converted to category)
    - April\_Cancellation: int64
    - May\_Cancellation: int64

 **Data Type Conversion**:

* **Converted 'Condition' Column to Categorical Type**:

**Justification**:

* Optimizes memory usage.
* Facilitates efficient grouping and statistical analysis.

**Filtering Data for May Analysis**:

* **Filtered Out Customers Who Canceled in April**:
  + - Ensures that the analysis of May cancellations only includes customers who were still subscribed at the beginning of May.

**Question 5: How many users in each condition canceled their plans in the month of receiving the email (April)? (1 pt)**

**Answer**:

* **Total Users in Each Condition: 500 users in Condition A and 500 users in Condition B.**
* **April Cancellations:**
  + **Condition A ($1 increase): 32 cancellations.**
  + **Condition B ($2 increase): 73 cancellations.**

**Visualization:**

* **Bar Chart:**
  + **The bar chart titled "April and May Cancellations by Condition" illustrates that Condition B had more than double the cancellations compared to Condition A in April.**

**Question 6: How many users in each condition canceled their plans in the month of the price increase (May)? (1 pt)**

* **Answer**:
* **Remaining Users After April**:
  + **Condition A**: 468 users (500 - 32 cancellations in April).
  + **Condition B**: 427 users (500 - 73 cancellations in April).
* **May Cancellations**:
  + **Condition A ($1 increase)**: **25 cancellations**.
  + **Condition B ($2 increase)**: **60 cancellations**.

**Visualization**:

* The same bar chart shows May cancellations, highlighting that Condition B again had more than double the cancellations compared to Condition A.

**Question 7: Run the proper analyses to test whether the two experimental conditions (A vs. B variants or conditions) led to higher cancellations during the month of receiving the email (April). Did Plan B lead to more cancellations during April compared to Plan A? Include p-values and counts or percentages. (5 pts)**

**Answer**:

**Statistical Test**: Chi-square test of independence.

**Contingency Table for April**:

|  |  |  |  |
| --- | --- | --- | --- |
| Condition | Not Canceled (0) | Canceled (1) | Total |
| Condition (A) | 468 | 32 | 500 |
| Condition (B) | 427 | 73 | 500 |

* **Chi-square Test Results:**
  + **Chi-square value: 15.29**
  + **Degrees of Freedom: 1**
  + **p-value: 0.000092**
* **Interpretation:**
  + **The p-value is far below 0.05, indicating a statistically significant difference in cancellations between the two conditions in April.**
  + **Plan B ($2 increase) led to significantly more cancellations in April compared to Plan A.**

**Visualization:**

* **Pie Chart for April Cancellations:**
  + **Shows that 69.5% of April cancellations were from Condition B, while 30.5% were from Condition A.**
* **Insights:**
  + **The pie chart visually emphasizes the disproportionate number of cancellations from Condition B.**

**Question 8: Run the proper analyses to test whether the two experimental conditions (A vs. B variants or conditions) led to higher cancellations during the month of the price increase (May). Did Plan B lead to more cancellations during May compared to Plan A? Include p-values and counts or percentages. (5 pts)**

* **Answer**:

**Statistical Test**: Chi-square test of independence (after filtering out April cancellations).

**Contingency Table for May**:

|  |  |  |  |
| --- | --- | --- | --- |
| Condition | Not Canceled (0) | Canceled (1) | Total |
| Condition (A) | 443 | 25 | 468 |
| Condition (B) | 367 | 60 | 427 |

**Chi-square Test Results**:

* + **Chi-square value**: 16.85
  + **Degrees of Freedom**: 1
  + **p-value**: 0.000040
* **Interpretation**:
  + The p-value is **well below 0.05**, confirming a **statistically significant difference** in cancellations between the two conditions in May.
  + **Plan B ($2 increase)** led to significantly more cancellations in May compared to Plan A.

**Visualization**:

* **Pie Chart for May Cancellations**:
  + Shows that **70.6%** of May cancellations were from Condition B, while **29.4%** were from Condition A.
* **Insights**:
  + The pie chart reinforces that Condition B continued to have a higher proportion of cancellations in May.
* **Line Chart for Cancellation Trends**:
  + Depicts the trend of cancellations over April and May for both conditions.
* **Insights**:
  + **Condition A** showed a slight decrease in cancellations from April to May.
  + **Condition B** also showed a decrease but maintained a higher number of cancellations overall.
  + The gap between the two lines highlights the consistent difference in cancellations due to the price increase amount.

**Question 9: Based on the analysis reported in QB-7 and QB-8, which of the two price increases do you recommend to the company? (1 pt)**

**Answer**:

 **Recommendation**:

* Based on the statistical analysis and visual evidence, I recommend implementing the **$1 price increase (Condition A)**.
* **Justification**:
  + **Lower Cancellation Rates**: Condition A consistently had significantly fewer cancellations in both April and May.
  + **Customer Retention**: The $1 increase minimized customer churn compared to the $2 increase.
* **Chi-square values**:
  + A **higher Chi-square value** indicates a greater difference between the observed and expected frequencies. In both April and May, the Chi-square values (15.29 and 16.85) suggest a significant deviation between the observed cancellation rates for Condition A ($1 increase) and Condition B ($2 increase).
* **p-values**:
  + The **p-values** for both April (0.0000921) and May (0.0000404) are **well below 0.05**, the commonly used significance level threshold.
  + A **p-value below 0.05** means there is strong evidence to **reject the null hypothesis**. In this case, you can conclude that there is a statistically significant difference in cancellations between Condition A and Condition B for both April and May.

**Conclusion:**

* For **both April and May**, the results suggest that the price increase of $2 (Condition B) led to **significantly more cancellations** compared to the $1 increase (Condition A).
* These p-values confirm that the difference in cancellations between the two conditions is **not due to random chance**.

 **Visual Support**:

* The bar chart and line chart visually demonstrate that Condition A led to fewer cancellations over time.
* The pie charts show a smaller proportion of cancellations attributed to Condition A in both months.

**Question 10: The company currently has 2 million users. The main subscription plan’s price was fixed at $9.99 throughout the past year. Based on the data from the previous year, the average monthly cancellation rate was .00002 per month (40 customers per month). Based on this information, would you recommend that the company increase the plan’s price to plan A or B, or do you recommend that they avoid making a price change altogether? (1 pt)**

**Answer**:

 **Analysis**:

* **Historical Cancellation Rate**: 40 customers per month out of 2 million users (0.002%).
* **Experimental Cancellation Rates**:
  + **Condition A**:
    - April: 32 cancellations (6.4%) out of 500 users.
    - May: 25 cancellations (5.34%) out of 468 users.
  + **Condition B**:
    - April: 73 cancellations (14.6%) out of 500 users.
    - May: 60 cancellations (14.05%) out of 427 users.

 **Recommendation**:

* **Implement the $1 Price Increase (Condition A)**.
  + **Rationale**:
    - While both price increases led to higher cancellation rates compared to the historical average, the $1 increase resulted in significantly fewer cancellations than the $2 increase.
    - The company needs to raise prices due to increasing operational costs, and the $1 increase balances revenue generation with customer retention.
* **Consider Communication Strategy**:
  + To mitigate cancellations, the company should enhance communication about the reasons for the price increase and highlight added value or upcoming features.

 **Visual Support**:

* The line chart shows that cancellations under Condition A decreased slightly from April to May, suggesting that initial reactions may subside over time.
* The visualizations collectively support the decision to choose the $1 increase over the $2 increase.

**Final Recommendation for StreamSmart Subscription Price Increase**

Based on the A/B test conducted to evaluate the impact of two price increase options on customer cancellations, the following insights and recommendations are derived:

**Key Findings:**

1. **April Cancellations**:
   * Condition A ($1 price increase) resulted in **32 cancellations**.
   * Condition B ($2 price increase) resulted in **73 cancellations**.
   * A Chi-square test confirmed that this difference is statistically significant (p-value = 0.000092), indicating that the $2 price increase led to significantly more cancellations before the actual price change was implemented.
2. **May Cancellations**:
   * After filtering out customers who canceled in April, Condition A ($1 increase) had **25 cancellations** in May.
   * Condition B ($2 increase) had **60 cancellations** in May.
   * A Chi-square test confirmed that this difference is statistically significant (p-value = 0.000040), showing that the $2 price increase caused more cancellations even after the price change was implemented.

**Conclusion:**

Both the April and May cancellation data show that the $2 price increase (Condition B) led to significantly higher cancellations compared to the $1 price increase (Condition A). Customers appear to have responded more negatively to the $2 price hike, with a much higher rate of cancellations both before and after the price change was implemented.

**Final Recommendation:**

Given the significantly higher cancellations under Condition B, I recommend **implementing the $1 price increase** (Condition A). This option resulted in fewer cancellations in both April and May, indicating better customer retention and a less negative impact on the subscriber base.

By adopting the $1 increase, the company can achieve a balance between raising revenue and minimizing the risk of losing subscribers.