**KPMG D&A Data Challenge Day**

1. **Introduction**

ООО «Радуга Мерчендайзинг» (JSC “Raduga Merchandising”) is a large face paint manufacturer based in Moscow, Russia. It was founded in the late 70’s by Alexander Popov, who at the time was a young man with a great passion for dressing up. Producing face paint started out as a mere hobby for Alexander and till this date he still cannot believe the great success of this company. Part of the success of his firm, according to Alexander, is related to the great popularity of face paint during sports tournaments. Especially in recent years, sales have never been higher as they are during global tournaments.

In order to make sure that his gut feeling of selling more face paint during FIFA World Cups, Alexander hired a data scientist, Alyona Sokolova. Alyona was eager to start her new job, as she saw many opportunities diving into the data of such a large corporation. The first thing she discovered is that Alexander’s gut feeling about the increase of sales during football tournaments was more than just a gut-feeling as it is in fact backed by data. Unfortunately, the data that she could get her hands on was not complete enough to make any real conclusions. As the 2018 FIFA World Cup is hosted in Russia, Alexander and Alyona approached KPMG with the question how to fully capitalise on data.

Due to the correlation between country flag colours and sales, we can state that if we can accurately predict the high performing teams and customize our colour mix to these teams, we can truly help our client.

Therefore, our question to your team is: *Can you make a model that predicts the outcome of the upcoming World Cup?*

For sake of simplicity, you may assume the following:

* The production costs per unit are equal for all colours and all countries
* The production capacity is large enough, so there are no limits on the production side
* Storage capacity however is limited, ordering too much of a specific colour could be more costly than running out of stock.
* Logistics costs to any 2018 World Cup cities are irrelevant.
* Alyona has already created a model that corrects for a country’s population, fan base, size of the economy, proximity to the tournament, etc. Therefore, the only thing she is interested in, is how she can include tournament performance in the model.

1. **Your task**

Your task is to give a 5 minute presentation to Alexander and Alyona about your solution to their problem. Your presentation is free format, but needs to take into account at least the following aspects:

* What data quality issues/measures did you take into account? What kind of controls do you identify?
* What approach did you take and which assumptions did you make during the analysis/building a model?
* Beware that you’re presenting to Alexander and Alyona, make sure to use good (and fancy) visualizations

1. **Outline of the day**

* 11:45 – 12:00: Introduction of Data & Analytics at KPMG
* 12:00 – 12:30: Introduction of the case + time for questions
* 12:30 – 13:30: Lunch (1 hour)
* 13:30 – 14:00: Meet your team, look at the data and prepare 90 second pitch
* 14:00 – 14:15: Every team pitches their strategy and the client will allocate additional budget to the teams based on these pitches
* 14:15 – 15:15: Work on case
* 15:15 – 15:20: Break
* 15:20 – 16:30: Work on case and prepare presentation
* 16:30 – 17:00: 5 minute presentation per team

1. **Useful information**
   1. **Data**

We already laid some groundwork for you and gathered some potentially relevant data. We think this data will suffice to create a reasonable estimation model, but you are free to enrich this data with data sources of your own choosing. If you choose to add data sources, please explain why these would be relevant and also why they are crucial to add under the given time constraints.

* + 1. **Participants of last and upcoming world cup:**

A list of all participating countries in the 2014 and 2018 world cup. The client already gathered data on the ratio of colours in the flags of these countries.

* + 1. **Results of previous world cups:**

We’ve provided a list of all matches and the results from the previous 5 world cups. This means we have data for the 1998, 2002, 2006, 2010 and 2014 tournaments. The data consists of 5 columns:

* year of match
* Home team
* Home team score
* Away team
* Away team score
  + 1. **FIFA ranking:**

In 1994, the FIFA first published a list of the ranking of all nations. Ever since, they have updated this list on a monthly basis. We’ve provided you with a list of the last ranking before each of the last 5 World Cups. The ranking in December 2017 has been added as well. This can be used as the ranking at the beginning of the 2018 World cup.

* + 1. **Rules:**

32 teams will participate in the 2018 World Cup. They will be divided over 8 groups of 4 teams. This normally happens based on their FIFA ranking, where the first team in each group is chosen from the organizing country and the 7 highest ranked teams. The next 8 teams on the ranking are in pot 2, and so on.

All teams in a group play one match against each other. Winning a match gives you 3 points in the group stage, a draw is 1 point for both teams and losing leads to 0 points. The top two teams of each group proceed to the knock-out stage. Teams that have the same number of points are ranked by the greater number of points obtained in the match between the concerned teams. If this was a draw, the teams are ranked on goal difference. When this is equal as well, the number of goals scored is the crucial number. In the scenario where this also doesn’t lead to a decision, a (fair) coin toss will make the decision.

* 1. **Considerations**

Although we don’t want to push you in any specific direction, we would like to give a few suggestions to keep in mind while working on the case.

* Think about checking your data quality. Bad data = bad analysis
* Before creating an analysis, think about the format and structure you wish to use.
* Always explain why you make specific assumptions/choices.
* Always test your models and, where possible, report significance of your outcomes
* Start with the end in mind. What/How do you want to present and what do you need to be able to do so. The board has a fancy for interactive clear visualizations.
* Build up your code in a modular way. All parts of the analysis can be written in smaller, more manageable chunks.

1. **Accelerating the process**

We have devised a system to help you in case you get stuck. Each team is awarded an initial budget of 75000 ₽ (Russian Rubles) at the start of the project. Sometimes, it is better to leverage the knowledge of external parties and/or the client to accelerate a project.

Therefore, we offer the possibility to ask for hints. These hints range from suggestions on an approach to buying pieces of the solution. The price for the separate parts is listed in brackets. If you already have part of the solution, you might receive a discount. This is however always up to the third party to decide.

During the initial phase, the client has additional budget to allocate to the idea he thinks has the highest chance of success. Therefore, you are asked to present your approach in a 90 second pitch. A good pitch might get you some additional budget to kick-start your project.

Obviously, the final deliverable and presentation will be judged by the client as well.

The following information is also present in the start-up Python script.

Exploration (15k)

* Look into the data what elements does it have (3k)
* What are the relevant variables for the simulation (5k)
* Can you find a function to describe the data? (7k)

Data cleaning (25k)

* You can choose to hire an external party to clean up your data. (25k)

Simulation (75k)

Knock-out phase (25k)

* single match with 50/50 chance for knock-out phase (7k)
* single round of the knock-out phase (6k)
  + make random pairs to match them up (lottery like) (5k)
* complete knock-out phase (7k)

group phase (25k)

* single match with 50/50 chance for group phase (5k)
* a way to keep track of the scores in the group phase (5k)
* make a round of the group phase (5k)
* complete simulation of the group phase (5k)
* how to extract the winner and follow-up in the group phase (5k)

weighting factor and final results (25k)

* introduce a weighting factor between the teams from external data (historical), for a suggestion see below (10k)
* combine group phase and knockout to a complete tournament (7k)
* make a way to keep track of tournament results (8k)

Colour distribution based on simulation results (10k)

* Develop a method to determine the colour distribution needed (10k)

Simulation results (100k)

* In case you get completely stuck, we offer you the possibility to buy our simulation result. This is considered a last resort in case you get completely stuck and will provide you the possibility to focus entirely on the other issues this client faces. It is however a very poor way to handle your resources.