Chapter 10

Stage 7: Reporting

10.1 GUERRILLA ANALYTICS WORKFLOW

Reporting is one of the types of analytics team outputs. It therefore happens at the end of the Guerrilla Analytics Workflow as illustrated in Figure 32. Given how late reporting happens in the workflow, any disruptions encountered at this stage can be very expensive. Mistakes in the underlying build, work products, or data introduce mistakes into the report. Refreshing data will require revisiting all analytics feeding into the report to make sure its analytics components are up to date, consistent, and correct.

10.2 WHAT IS A REPORT?

The previous chapter discussed how to produce work products and how this is something the Guerrilla Analytics team does from the very beginning of the project. Since much effort is devoted to exploring and understanding the data in collaboration with business users and customers, you would expect there to be many work products and iterations of those work products. This is the case and several chapters of this book have been devoted to coping with the disruptions of this highly iterative and dynamic cycle.

On many projects, there will be occasions where a weightier deliverable also needs to be produced. This is typically a written document or presentation that is communicating key project findings to project stakeholders and sponsors. The document is often produced at milestones in a project or to summarize a project's findings at the end of the project. The format is predominantly text, which is then interleaved with numbers, figures, tables, and analyses. Some of these will have been produced by the Guerrilla Analytics team but it is possible that others will have been produced by members of the broader project team. Here are some examples.

- In a forensic investigation, a report is a key deliverable that details the results of the investigation. Its contents may be presented in a court of law or used in a firm's internal HR process. The report may draw heavily on data "evidence" produced by the analytics team.
- In management consulting projects, the report may deliver key recommendations for a customer or shareholder. Examples include decisions on a firm's strategy, a decision to pursue a merger, the launch of a new service, etc.

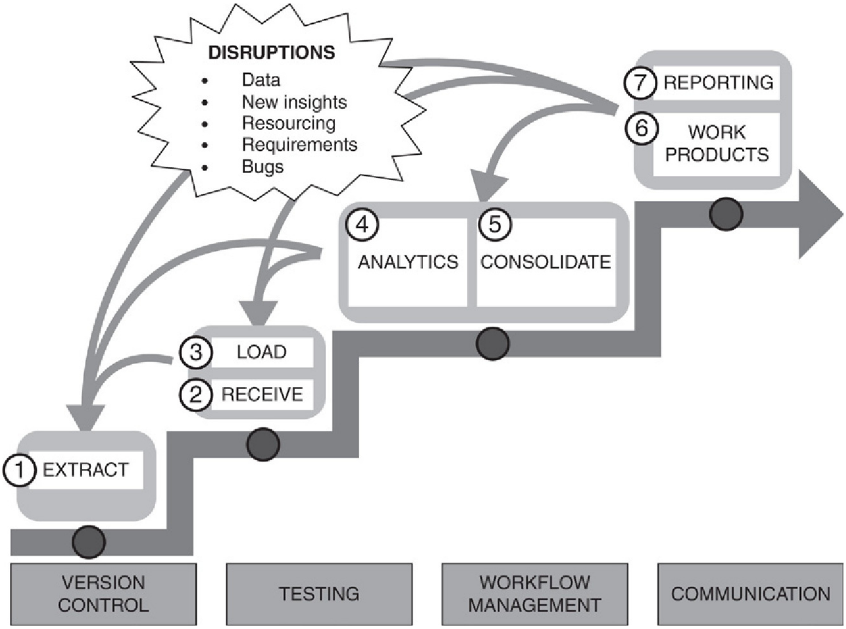


FIGURE 32 **The Guerrilla Analytics workflow**

- In a transformation project, the report may summarize current customers, potential for churn and recommend a new strategy, product and marketing to counteract this churn. These report recommendations could feed into significant investment decisions for the business.
- A banking conduct risk project will describe the types of risk models created and the assumptions and data on which these models were built. Many scenarios may be modeled and commented on with recommendations for operational changes in the bank.
- A well written scientific paper must be reproducible with clear conclusions for it to pass peer review and be published.
- A well-archived Guerrilla Analytics project will have a closing document that describes the data analyzed, how it was sourced, assumptions made, analyses, and recommendations.

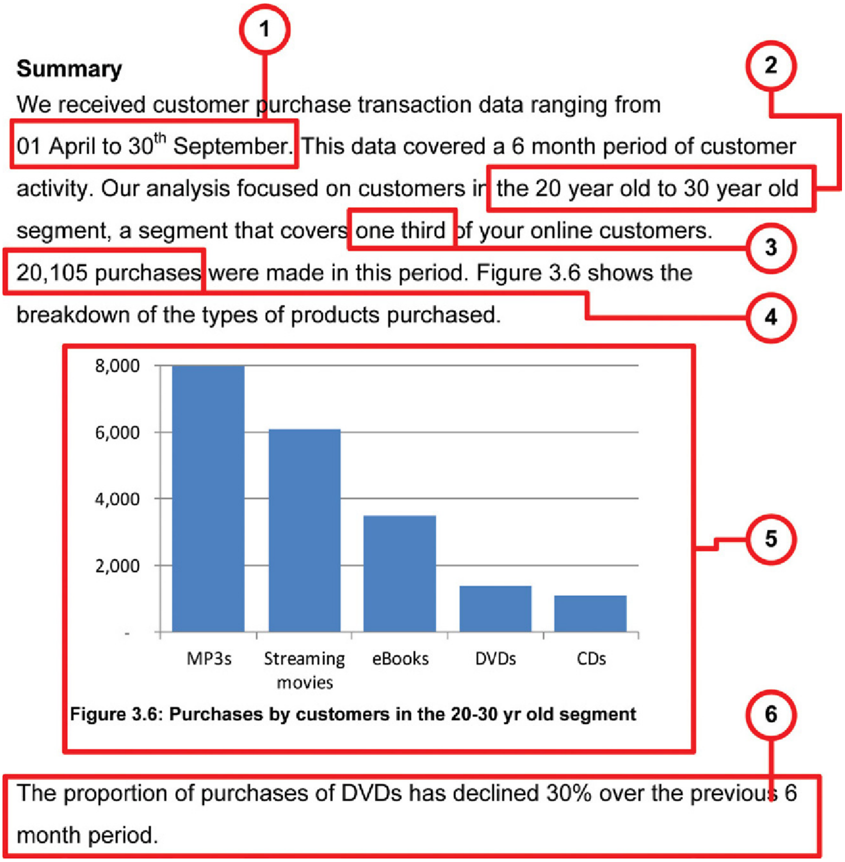
The list of reporting examples goes on but hopefully you have the idea. Work products are the day to day evolving analytics that represent the team and customer's growing understanding of the data. Reports are set-in-stone work products that absolutely have to be correct, traceable, tested, etc. Once delivered, there is little scope for retraction, iteration, and corrections.

Reports have other significant inputs beyond analytics work products. Writing a report means combining some form of write-up and commentary with analytics results from the team. These results might be a single number embedded in a larger paragraph of text. Results might also be relatively clear-cut figures or tables taken directly from data analytics work products. A more difficult scenario is when the analytics team's result goes to another team member or a business analyst who then modifies the work product further before it contributes to the report. Clearly this presents a particularly difficult challenge for the traceability and data provenance of the Guerrilla Analytics team's outputs.

10.3 WHY REPORTS ARE COMPLICATED

There are some clear risks in report writing in addition to the general risks around work products that were discussed in the previous chapter. Let us first look at a typical report example.

Figure 33 shows a page from a report. This is a completely fabricated example and is deliberately simplistic. However, it has many of the features you see in a typical report and therefore nicely illustrates many of the sources of lost data provenance for the Guerrilla Analytics team.



Summary

We received customer purchase transaction data ranging from 01 April to 30th September. This data covered a 6 month period of customer activity. Our analysis focused on customers in the 20 year old to 30 year old segment, a segment that covers one third of your online customers. 20,105 purchases were made in this period. Figure 3.6 shows the breakdown of the types of products purchased.

Figure 3.6: Purchases by customers in the 20-30 yr old segment

The proportion of purchases of DVDs has declined 30% over the previous 6 month period.

FIGURE 33 An example report

The report contains a small paragraph of text. Several numbers are mentioned in this paragraph. There is also a simple figure, which is referenced from the paragraph. Such a report presents a number of challenges for traceability as highlighted in Figure 33.

- **A date range for the data is quoted.** Who sourced this date range and is it correct? Is there a particular date field in the data that could reproduce this date range?
- **A customer segment is quoted.** Again who sourced this and is it correct? How was this segment determined from the data? Was it already in the data or was some segmentation performed?
- **It is stated that this customer segment is one-third of the total customer population.** Was this calculated from the same data used in the previous quotes? Do the authors mean one-third in numbers or in sales value or by some other metric?
- **A total number of purchases for the customer segment in this date range are quoted.** Again who calculated this total? Are we sure it was calculated from the exact same data as provided by the other quoted numbers?

Even if the above challenges on accurately specifying data populations are overcome, there remain challenges around consistency of components of the report.

- **There is a figure summarizing purchase types.** Do the totals and breakdown in this figure tie out to the numbers quoted in the preceding paragraph?
- **The data are compared to a previous date range to make a claim about a decline in sales.** Where was this number sourced from or is it an anecdotal reference?

This simple example illustrates how one paragraph of text and a summary figure can generate many questions around data provenance, correctness, and consistency.

10.4 REPORT COMPONENTS

It is helpful when discussing reports to think if them as having the components illustrated in Figure 34. The components are divided into written components and analytical components.

- **Written components:** These are the concern of the report authors. They consist of the usual report sections such as scope, introduction, comments, conclusions, recommendations that are part of any analytical report. They are not directly dependent on data.
- **Analytical components:** These are the concern of any teams doing analysis work, including the Guerrilla Analytics team. The types of analytical components are figures, tables of data, text that references these figures and tables, and text that references data in general.

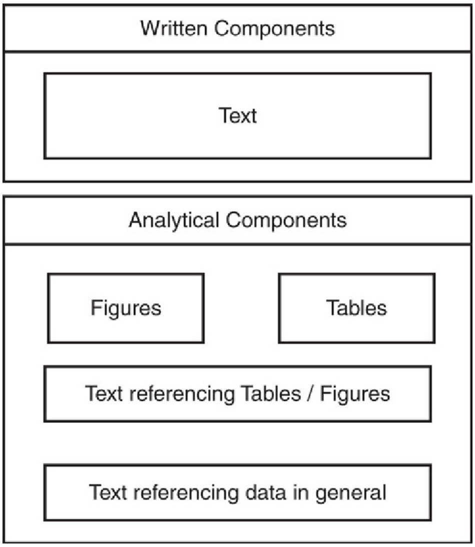

Written Components
Text
Analytical Components
Figures
Tables
Text referencing Tables / Figures
Text referencing data in general

FIGURE 34 **Components of a report**

Have a look through the example report at the start of this chapter and see if you can identify each of these components.

10.5 PITFALLS AND RISKS

Thinking about the questions raised by the sample report of the earlier section, we can identify the following risks for the Guerrilla Analytics team. Broadly these fall into two categories.

10.5.1 Data Provenance

These risks relate to recognizing a particular analytics component in a report and being able to trace its origins.

- **Data modified after leaving the team:** The data that leaves the analytics team can be further modified by somebody outside the team before it enters the report. This is done for two reasons: the data needs to be modified for presentation purposes or somebody wants to do their own analysis before putting a final work product into the report. The latter could happen when an analytical domain expert such as an insurance actuary needs to perform his/her own analysis.
- **A data team's work products cannot be identified:** Not all numbers in a report will have come from the analytics team. In bigger projects there may be business analysts, accountants, scientists, and other domain experts who are perfectly capable of generating their own data and outputs. Alternatively, a single number quoted deep in some paragraph may be a complex analytics team work product of several code files that evolved over several versions.

10.5.2 Consistency

Consistency relates to the risk of components of the report not agreeing with one another.

- **Consistency with the analytics team:** Are the components used in the report consistent with what the analytics team would produce?
- **Self-consistency:** Are components within the report consistent with one another? Components can become inconsistent if they come from separate sources or are modified in inconsistent ways. In our illustrative example, the figure may have come from the analytics team while the summary comments on the figure may have been written by a different team and calculated from an analytics data sample. They all involved the same data, however, so they had better all be consistent with one another.

10.5.3 Implications

These risks have several implications for the Guerrilla Analytics team.

- First, without a proper process in place there can be disagreement over which components came from the analytics team and which came from some other part of the project.
- Second, if a component cannot be identified, there is no way to reproduce the given component.
- Third, and perhaps most seriously, there is a lot of scope for an incorrect report.

Let us now look at the key ways to mitigate these risks.

War Story 13: The Quick Fix

Claire was managing the analytics team working on a project which involved writing formal reports for a client every 6–8 weeks. The report writers would hide themselves away for a week, writing up their findings with data they had received from Claire's analytics team in a variety of work products over the course of the previous 6 weeks. Timelines were always extremely tight and all of this happened without bringing analytics onto the reporting team. The report would be reviewed internally and then published. This was how things went until somebody raised the question of the provenance of all numbers used in the reports. Panic!

The quick fix was to revisit all possible analytics components in the report and check that they could be reproduced. A dump of figures, spreadsheets, and data came back to Claire's team with a copy of the published report. Some of the data contained recognizable analytics work products (Claire was using a work product UID in the team's file names). Some of the data seemed to have been modified or even had nothing to do with the analytics team at all.

Claire had to sideline several analytics team members for a week as they went through the 60 page report writing analytics queries that would reproduce report numbers and highlighting numbers that did not involve her team at all.

Imagine how painful this was. Think of the number of ways to filter and pivot a dataset to turn it into a report table. Think of having to reproduce a single number quoted in the middle of a paragraph of text with little or no analytics context. Much effort could have been saved by involving Claire's analytics team with the report writers at the time of writing and agreeing a process for maintaining data provenance in the report.

10.6 PRACTICE TIP 42: LIAISE WITH REPORT WRITERS

10.6.1 Guerrilla Analytics Environment

Very often, the person writing the report is not from the analytics team. There may be more than one writer and some or none of those people may be from the analytics team. How can you maintain data provenance if you do not have visibility of how analytics work products are being used?

10.6.2 Guerrilla Analytics Approach

Whenever a report is being produced, an analytics team member should be part of the report writing team. Their job is to liaise with the report writing team and do the following key things.

- Have an overview of the report content.
- Identify when new analytics work products are required and when existing work products are being used.
- Ensure that every analytics work product used in the report can be traced back to its work product UID.

10.6.3 Advantages

The main advantage of a mixed team of Guerrilla Analytics and report writers is that data provenance of report content is maintained. If this tip is applied successfully, every analytics output in a report should be identifiable by a work product UID. This ensures full reproducibility of the report content.

10.7 PRACTICE TIP 43: CREATE ONE WORK PRODUCT PER REPORT COMPONENT

10.7.1 Guerrilla Analytics Environment

We know that some of the Guerrilla Analytics team's outputs will appear in a report as an individual number, a statement about another component, a table or a figure. It is difficult for the analytics team to know in advance the context in which their work product will be used. It would therefore be very useful to be able to easily reproduce any given analytical component that appears in a report. You would rather avoid having to do an excessive amount of searching through a history of analytics work product, some of which may have to be reworked.

10.7.2 Guerrilla Analytics Approach

Every report component from the Guerrilla Analytics team should be a work product with its own work product UID. This should happen regardless of how simple or complex the component in the report is.

Consider a simple example where the report must have a sentence such as "Data was analysed from April to September." "April to September" should be a simple work product with its own UID. In practice this may be a one line query to calculate the earliest and latest month from the data. Nevertheless, there would now be a definitive traceable record of the date field and dataset from which the quoted date range was generated.

10.7.3 Advantages

The advantages of this tip are as follows.

- **Data Provenance:** Any analytics component of the report can be identified and reproduced quickly without rework or reinterpretation.
- **Efficiency:** Time is saved in trying to ensure that a report has the latest numbers and that those numbers are consistent with one another. Every report component from the analytics team has full data provenance so you know how it was produced in code and the version of the data it was based on.
- **Explaining Changing Numbers:** Because report components are based on actual work products, they have an identifiable version. Future project reports can explain why their quoted numbers and analyses may differ from previous reports.

10.8 PRACTICE TIP 44: MAKE PRESENTATION QUALITY WORK PRODUCTS

10.8.1 Guerrilla Analytics Environment

The Guerrilla Analytics team will often produce work products that are further modified by other team members or the report writers themselves. For example, a fairly raw dataset may be filtered or aggregated to produce a summary table appropriate for a report. The more an analytics component is modified outside the analytics environment, the more difficult it is to reproduce it at a later date and so these types of modifications need to be discouraged somehow.

10.8.2 Guerrilla Analytics Approach

Analytics work products for a report component should be tailored to be as close as possible to the format required in the report. For example, if a report table should have certain headings and row labels, then you should produce those headings and row labels in the analytics work product rather than leaving it to the report writer. If a figure should have certain axis labels and limits then you should set those axis limits and labels in the work product code, not after the figure has been handed over for inclusion in the report.

10.8.3 Advantages

This tip eliminates the need for report writers or nonanalytics team members to modify work products and break data provenance. As far as possible, if they have everything they need in the format in which they need it then everybody's life will be a lot easier.

10.9 EXTREME REPORTING

You may be reading this chapter and thinking it all seems a little brittle. Have an analyst sit with report authors identifying work products? Make work products presentation ready so a link isn't broken between the analytics team and its work products? Surely there is a better toolset for doing this? There are some options worth mentioning.

10.9.1 In-Line Documentation

Much of what was discussed in this chapter involved simple processes for bringing the analytics into the report. That is, a report document exists in a word processor application and we want to embed analytics work products into the word processor document.

In-line documentation turns this approach on its head and takes the report into the analytics. A large "code" file is written that generates all the required outputs such as numbers, tables, and figures. This same code file contains the text for the report. When in-line documentation tools are run on this "code file" it typesets the text and generates the latest versions of the analytics components to produce a final document. This is the approach taken by tools such as Sweave (Leisch, 2002).

There are some challenges to this approach when used in a Guerrilla Analytics project.

- **Programming language variety:** The reality is that Guerrilla Analytics involves analyses with a wide variety of tools and datasets. There is no one size fits all analytics language that covers all the team's needs.
- **Need for iterative reviews:** Many of the team's analyses are iterative and require skilled interpretation before they can be used. You cannot just click "go" and trust that the outputs are ready to go straight into a report.
- **Incompatibility with office software:** Many report writers prefer to use office suites that do not look like "program code."

So unfortunately, while in-line documentation seems like a promising approach, the opportunity to use it in Guerrilla Analytics projects is rare. Only projects with short analytics execution times and a small number of report writers are appropriate.

10.9.2 A Simple Alternative

A simple alternative is to provide a word processor add-in with the following functionality.

- **Tagging:** Can tag analytics components (numbers, text, figures, and tables) with a work product UID.
- **Listing:** Can create a report listing all work product UIDs used in a document.

Then the only change to a team's approach to report writing would be to teach report writers to tag all analytics components with the work product UID they received from the analytics team. When a report needs to be reproduced, the relevant work products UIDs can be easily listed.

10.10 WRAP UP

This chapter has discussed report writing in Guerrilla Analytics projects. You have learned.

- **What a report is:** A report is a formal document that is a combination of written content and analytics work products. Unlike a typical work product that may be iterative and collaborative, a report is generally a complex document perhaps with multiple authors and many components both from analytics and from other teams.
- **Report components:** Regardless of how complex or lengthy a report is, it fundamentally consists of written components and analytical components. The analytical components are tables, figures, text that references both tables and figures and finally, text that references data in general.
- **Risks in reporting:** The nature of reporting where the report writers often take analytics outputs for inclusion in the report raises the following risks.
  - Data modified after leaving the team.
  - The team's work products cannot be identified.
  - Consistency of report components with the analytics team outputs.
  - Consistency of report components with one another within the report.
- Simple practice tips mitigate the risks in reporting.
  - Stay close to the report writers and liaise with them.
  - One work product per report component.
  - Make presentation quality work products.